UNIVERSIDAD POLITÉCNICA DE MADRID

**ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN**

**ETSIT
UPM**

*Design and Evaluation of Statistical Parametric
Techniques in Expressive Text-To-Speech:
Emotion and Speaking Styles Transplantation*

TESIS DOCTORAL

Jaime Lorenzo Trueba

Máster en Ingeniería de Sistemas Electrónicos

2016

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

**ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN**

**ETSIT
UPM**

*Design and Evaluation of Statistical Parametric
Techniques in Expressive Text-To-Speech:
Emotion and Speaking Styles Transplantation*

TESIS DOCTORAL

Jaime Lorenzo Trueba

Máster en Ingeniería de Sistemas Electrónicos

Supervisor: Juan M. Montero Martínez

Doctor Ingeniero de Telecomunicación

Supervisor: Roberto Barra Chicote

Doctor Ingeniero de Telecomunicación

2016

UNIVERSIDAD POLITÉCNICA DE MADRID
DEPTO. DE INGENIERÍA ELECTRÓNICA
E.T.S.I de TELECOMUNICACIÓN

TESIS DOCTORAL

TÍTULO: Design and Evaluation of Statistical Parametric Techniques in Expressive Text-To-Speech: Emotion and Speaking Styles Transplantation.

DOCTORANDO: D. Jaime Lorenzo Trueba.

DIRECTORES: D. Juan Manuel Montero Martínez, D. Roberto Barra Chicote.

El día _____ de _____ de _____, el Magnífico y Excelentísimo Señor Rector de la Universidad Politécnica de Madrid, designó el siguiente Tribunal para evaluar la Tesis Doctoral arriba mencionada:

Presidente: Luis A. Hernández Gómez

Secretario: Ricardo de Córdoba Herralde

Vocal 1º: Javier Macías Guarasa

Vocal 2º: Simon King

Vocal 3º: Ignacio Iriondo Sanz

Suplente 1º: Eduardo Rodríguez Banda

Suplente 2º: David Escudero Mancebo

Realizado el acto de Lectura y Defensa de dicha Tesis Doctoral en Madrid, el día _____ de _____ de _____, en la ETSI de Telecomunicación, el Tribunal acuerda otorgarle la siguiente calificación:

CALIFICACIÓN: _____

EL PRESIDENTE

EL SECRETARIO

LOS VOCALES

A ti, mi compaera de viaje

Agradecimientos

Quiero empezar este documento agradeciendo a todas y cada una de las personas que me han ayudado a que esta tarea llegue a su final, empezando por mis dos directores: Juancho y Roberto. Sin vosotros esto no habría sido posible desde un principio. Quiero agradecer también a mi director de estancia y no solo de estancia, Junichi. Porque no sólo has contribuido al desarrollo de mi tesis sino también a darle un futuro.

También quiero agradecer su apoyo al resto de integrantes del Grupo de Tecnología del Habla: Manolo, Ricardo, Luisfer (¡en la distancia!), Fernando y Rubén (gracias a ti empecé esto) y a mis compañeros: Bea, Julián, Christian, Vero, Syah... Por supuesto también a Luis Hernández y Javier por ayudar en las fases preliminares de este final de tesis a que el resultado final fuera el que hoy en día es.

Por supuesto también quiero agradecer a todos los integrantes de Simple4All, ARABOT, IN-APRA. Todos habéis formado parte de un grupo de trabajo que me ha permitido desarrollar mis capacidades en un entorno de trabajo y de investigación inmejorable. También a mi universidad, la Universidad Politécnica de Madrid y al Departamento de Ingeniería Electrónica por darme los medios para sobrevivir y desarrollarme durante todo este tiempo.

No puedo olvidarme tampoco de todos los que habéis participado con infinita paciencia en mis evaluaciones, tanto familia (¡todos vosotros!) como amigos (también todos, para que engañarnos). Soy muy consciente que ha sido muchas veces duro y aburrido y definitivamente sin vosotros no habría sido posible.

A Chema, Sheila, Fonta, Andrea, Varo, Giselle, Álvaro, Marta... Habéis ayudado a que los días difíciles fueran más fáciles (además de tener una paciencia infinita con mis evaluaciones, ¡nunca podré agradecerlo lo suficiente!).

A mis padres Raquel y Jorge y a mi hermano, Jorge. Porque la familia siempre está ahí, y yo no sería la persona que soy sin vuestra influencia a lo largo de mi vida.

Por último a Marina, mi compañera de vida desde prácticamente el principio de este viaje. Juntos hemos crecido, vivido y experimentado. Sin tu apoyo y tu amor no sé cómo hubiera sido posible nada de esto. Espero con ilusión nuestro futuro.

En conclusión, gracias a todos por todo. Sin vosotros nada hubiera sido igual.

Resumen

Acompañando al incremento en la implementación de las tecnologías en la vida cotidiana ha habido una gran concienciación sobre la utilidad de las tecnologías de habla como nexo entre humano y máquina. En concreto en el campo de la conversión texto a voz ha habido grandes avances en naturalidad y calidad aupados por las fuertes inversiones realizadas por grandes multinacionales tecnológicas. La mayor parte de estas mejoras se deben a los avances en las técnicas basadas en redes neuronales profundas, capaces de generar voces de muy alta calidad en aplicaciones tradicionales. A pesar de ello, conseguir sistemas de síntesis de voz expresivos de alta naturalidad sigue siendo un problema debido a la escasez de datos de entrenamiento. Generalmente conseguir buenos sintetizadores requiere elevadas cantidades de información, algo fácil de conseguir en habla leída pero mucho más difícil si pensamos en expresividades concretas como emociones o estilos de habla. Tradicionalmente la forma de solucionar esta limitación ha sido la aplicación de técnicas de adaptación para convertir modelos robustos entrenados con datos de diferentes fuentes en el locutor y expresividad deseada, pero esto también tiene sus limitaciones.

Esta tesis busca solucionar los problemas existentes en las técnicas tradicionales de síntesis de voz expresiva centrándose en un aspecto concreto: la escalabilidad. Por una parte, conseguir datos para todas las expresividades y locutores es una tarea inabarcable, y por otro lado las soluciones tradicionales suelen introducir otra clase de limitaciones en la naturalidad o intensidad expresiva. Esta tesis presenta en primer lugar un sistema capaz de **controlar las características paralingüísticas de la voz** (los cambios dinámicos en la forma de hablar que transmiten información expresiva), permitiéndonos controlar la intensidad expresiva generada en la síntesis. Así se vuelve posible generar diferentes niveles de intensidad expresiva partiendo de un único nivel de partida. En segundo lugar presenta un sistema de **trasplante de información paralingüística** que permite incluir información expresiva en un locutor destino sin necesidad de tener información expresiva de dicho locutor. Finalmente se introduce un **sistema completo de conversión texto a voz expresivo automático** capaz de detectar qué expresividad es la más adecuada para transmitir el texto que se desea sintetizar, y posteriormente generar el mensaje hablado correspondiente sin ninguna interacción por parte del usuario más allá de introducir el texto deseado.

Todos los sistemas propuestos y sus variantes han sido evaluados perceptualmente gracias a la colaboración de numerosos voluntarios, lo que nos ha permitido certificar los beneficios que aportan a las técnicas de síntesis de habla expresiva. Las evaluaciones fueron diseñadas para

minimizar posibles sesgos en las evaluaciones y fueron analizadas con estrictos criterios de significancia estadística, fijando bandas de confianza del 95%. Los resultados de las evaluaciones demostraron que el sistema de control de características paralingüísticas es capaz de controlar de forma prácticamente lineal la intensidad emocional generada (0,97 de correlación con el sistema ideal) a la vez que aporta un 70,6% de tasas de reconocimiento de las emociones sintetizadas, valores muy superiores a los conseguidos con las técnicas tradicionales. En cuanto al sistema de trasplante de información paralingüística vemos que es capaz de mejorar en 1,2 puntos de la escala MOS la intensidad emocional generada en el locutor destino a costa de únicamente 0,4 puntos de calidad, traducándose en una media de 87% de preferencia de los evaluadores hacia un sistema trasplantado frente a un sistema no emocional. Los resultados se confirmaron también para el trasplante de estilos de habla como retransmisión de partidos de fútbol o discurso político. Por último se probó las capacidades de trasplante de emociones a través de idiomas, de emociones de un locutor español a uno inglés concretamente, dando resultados razonables.

Finalmente, para la evaluación y diseño del sistema completo de conversión texto a voz expresivo automático empezamos por grabar una base de datos de estilos de habla de un único locutor en castellano, Spanish Speaking Styles. Esta base de datos es la primera de su clase pensada explícitamente para la tarea de síntesis de estilos de habla, tarea para la que constar con un único locutor aumenta significativamente la calidad. Con estos datos se evaluaron tres enfoques del sistema de síntesis: uno discreto en el que se predice el estilo de habla a partir del texto y se sintetiza directamente, uno semi-continuo en el que además se extrae información del texto que permite generar una intensidad expresiva más adecuada y finalmente un sistema continuo que no depende de ninguna predicción sino que directamente utiliza información léxica para generar la voz. Los resultados muestran que los tres sistemas son capaces de mejorar significativamente los resultados de la síntesis tradicional en términos de adecuación de la voz generada (en media entre 1,0 y 1,5 por encima) y en intensidad expresiva (entre 1,1 y 2,0 mejor dependiendo del sistema). En cuanto a calidad los sistemas discreto y semi-continuo no sufren ninguna caída en comparación a los mejores sistemas neutros (4 puntos en la escala MOS en media). El sistema continuo, principalmente debido a la naturaleza semi-continua de los datos de entrenamiento, introduce artefactos que estropean la calidad de la voz generada.

Abstract

Together with the rise in the implementation of technology in daily life the awareness of the usefulness of speech technology as a nexus between humans and machines has risen too. Concretely in the field of text to speech conversion there have been huge advancements in terms of naturalness and quality spurred by big investments carried out by some of the most successful multinational technological companies. Most of the advances have come thanks to the progress made in techniques based on deep neural networks, capable of generating voices of very high quality in traditional applications. Nonetheless, designing expressive text to speech synthesis systems with high naturalness is still a problem mainly because of data scarcity. Generally obtaining good synthesizers require vast amounts of information, something that is reasonably easy to obtain for neutral, read speech, but significantly harder to obtain when thinking about particular expressiveness such as emotions or speaking styles. Traditionally, the way to solve this limitation has been the application of adaptation-based techniques that convert robust models trained with diverse and abundant data into models for the desired speaker and expressiveness, but those approaches also have their limits.

The present document aims to solve some of the existing problems in the traditional expressive speech synthesis techniques by focusing in one particular aspect: scalability. On one hand, gathering data for every expressiveness and speaker is a task that cannot realistically be tackled. On the other hand, traditional solutions typically induce limitations in naturalness or expressive intensity. This thesis presents first of all a system capable of **controlling the paralinguistic properties of speech** (the dynamic changes in the way of speaking that convey expressive information), allowing us to control the expressive strength produced by the synthesis. Thus, it becomes possible to generate different levels of expressive intensity even if the source data does not contemplate it. Secondly, it presents a system for **transplanting the paralinguistic information into a different target speaker** which allows us to imbue expressive information into a target speaker for whom no expressive information is available. Finally it introduces a **complete and automatic expressive text to speech system**, capable of detecting the most adequate expressiveness for conveying the text to be synthesized and then generate the corresponding audio file without any further interaction from the user.

All the proposed systems and their variants have been perceptually evaluated thanks to the collaboration of numerous volunteers, which has allowed us to certify the benefits that our systems

provide to the field of expressive speech synthesis. The evaluations were designed to minimize evaluation bias and were analyzed with strict statistical significance criteria, fixing a 95% confidence band. The results of the evaluations proved that the paralinguistic features control system is capable of controlling in a practically linear fashion the produced emotional strength (correlation coefficient of 0.97 when compared with the ideal linear system) while providing a 70.6% of emotion identification rates. These results are significantly better to those achievable by traditional techniques. In terms of the transplantation system, we see how it is capable of increasing by 1.2 points in the MOS scale the produced expressive strength of the target speaker when compared to the non transplanted version at a cost of just 0.4 points in speech quality. This translates into an average preference rate for the transplanted system of 87% when compared with a neutral system. The results are consistent when transplanting speaking styles such as live sports broadcasts or political speech. The system was also modified to add cross-lingual transplantation capabilities, which worked reasonably well when transplanting Spanish paralinguistic information into an English speaker.

Finally, for the evaluation and design of the complete expressive text to speech system we began by recording a single speaker Spanish speaking styles database, SSS. This database is the first of its kind to be designed explicitly with speech synthesis in mind, a task for which using data for a single speaker significantly helps increase quality. With these data we evaluated three approaches to the speech synthesis system: a discrete one in which the most adequate speaking style is predicted from the genre of the text to be synthesized, a semi-continuous approach that improves on the discrete system by adding expressive intensity information and finally a continuous system that does not rely on gender prediction techniques and directly makes use of lexical information to generate the speech. The results show how all three systems are capable of significantly improving traditional speech synthesis systems in terms of adequacy (average improvements between 1.0 and 1.5 depending on the system) and in terms of produced expressive strength (improvements between 1.1 and 2.0 in this case). In terms of speech quality both the discrete and the semi-continuous system are capable of maintaining neutral speech synthesis rates (averages of 4 points in the MOS scale). The continuous system, mainly due to the semi-continuous nature of the training data, introduces artifacts in the synthetic speech that harm the produced speech quality.

Glossary

AVM Average Voice Model

CAT Cluster Adaptive Training

CSMAPLR Constrained Structural Maximum A Posteriori Linear Regression

CVSM Continuous Vector Space Modeling

DNN Deep Neural Network

F0 Fundamental Frequency

HMM Hidden Markov Model

HSMM Hidden Semi-Markov Model

LSA Latent Semantic Analysis

MCEP Mel Cepstral Coefficient

ML Maximum Likelihood

MOS Mean Opinion Score

MR-HMM Multi-Regression HMM

NAV Neutral Average Voice

RNN Recurrent Neural Network

SAT Speaker Adaptive Training

SEV Spanish Expressive Voices

SPSS Statistical Parametric Speech Synthesis

SSS Spanish Speaking Styles

SVD Singular Value Decomposition

TDM Term-Document Matrix

TTS Text-To-Speech

US Unit Selection

VCTK Voice Cloning ToolKit

WTDM Weighted TDM

Contents

Glossary	ii
1 Introduction	1
1.1 Motivation	1
1.2 Organization	3
1.3 Objectives	3
1.3.1 Design of a New Method for Controlling the Strength of Paralinguistic Properties of Speech	4
1.3.2 Design of a New Method for Transplanting Paralinguistic Information between Speakers	4
1.3.3 Design of an Expressive TTS System: From Text Genre to Speaking Styles	5
2 State of the Art	7
2.1 Expressive Speech	7
2.1.1 Emotions	7
2.1.2 Speaking Styles	9
2.2 Text to Speech Synthesis	11
2.2.1 Introduction	11
2.2.2 Unit Selection Based TTS	12
2.2.3 Statistical Parametric Speech Synthesis	13
2.2.4 Strengths and Weaknesses of the Techniques	16
2.2.5 Neural Network based synthesis	17
2.3 HMM-Based Synthesis Adaptability Techniques	19
2.3.1 Interpolation Techniques	19
2.3.2 Adaptation Techniques	20
2.3.3 Multi-Regression HMM adaptation	23

2.4	Genre Classification	24
2.4.1	Introduction	24
2.4.2	Document representation and weighting schemes	25
2.4.3	Discrete genre modeling	28
2.4.4	Continuous vector space modeling	29
2.5	Expressive TTS	31
2.5.1	Introduction	31
2.5.2	Rule-based vs. Corpus-based Expressive TTS	32
2.5.3	Speaker Dependent vs. Speaker Independent Techniques	33
2.5.4	Expressiveness Transplantation Methodologies	34
2.5.5	Cross-language Expressive TTS	35
2.6	Expressive Corpora	36
2.6.1	Emotional Corpora	36
2.6.2	Speaking Styles Corpora	36
2.7	Evaluation Methods in TTS	38
2.7.1	Latin Square	38
2.7.2	Perceptual Evaluations	39
2.7.3	Statistical Significance	40
2.7.4	Proposed Evaluation Framework	41
3	Controlling the Strength of Paralinguistic Properties of Speech	43
3.1	Introduction	43
3.2	Adaptation based paralinguistic features control system	43
3.2.1	Control through adaptation	44
3.2.2	Proposed System	45
3.3	Perceptual Evaluation	46
3.3.1	Evaluation Design	46
3.3.2	Emotion Identification Rates	47
3.3.3	Speech Quality	50
3.3.4	Emotional Strength	51
3.3.5	Iberspeech 2012 Evaluation	53
3.3.6	Controlling Emotions in a Spoken Conversational Agent	54
3.4	Paralinguistic Features Control System Conclusions	55
4	Transplanting Paralinguistic Information between Speakers	56

4.1	Introduction	56
4.2	Proposed system	56
4.2.1	Emotion Transplantation based on Cascade Adaptation	57
4.2.2	Proposed Emotion Transplantation Method	58
4.3	Evaluation	60
4.3.1	Preliminary Evaluations	60
4.3.2	Evaluations Design	62
4.3.3	Evaluation 1: average emotion transplantation system vs. neutral system .	64
4.3.4	Evaluation 2: proposed emotion transplantation system vs. neutral system	65
4.3.5	Evaluation 3: proposed emotion transplantation system vs. average emo- tion transplantation system	68
4.4	Evaluation of the transplantation system in a real robotic application	69
4.4.1	Robot description	69
4.4.2	Robot evaluation design	70
4.4.3	Robot evaluation results	71
4.5	Paralinguistic Information Transplantation System Conclusions	73
5	Expressive TTS System: From Text Genre to Speaking Styles	75
5.1	Introduction	75
5.2	Spanish Speaking Styles	75
5.2.1	Database Design	76
5.2.2	Recording Environment	77
5.3	Prediction based TTS systems	78
5.3.1	Centroid based predictor	78
5.3.2	Discrete approach	79
5.3.3	Semi-continuous approach	80
5.3.4	Perceptual Evaluation	83
5.4	Continuous system	90
5.4.1	Proposed System	91
5.4.2	Perceptual Evaluation	92
5.5	Systems Comparison	94
5.6	Expressive TTS System Conclusions	97
6	Conclusions	99
6.1	Future Work	100

7	Contributions	102
7.1	Refereed Journal Papers	102
7.2	Book Chapters	102
7.3	Refereed Conference and Workshop Papers	102
7.4	Other Contributions	104
A	Other Studies on Transplantation	126
A.1	Preliminary Speaking Styles Transplantation Evaluation	126
A.1.1	Alternatives to Transplantation: Copying the Speaking Style Average Model	126
A.1.2	Perceptual Test Description	127
A.1.3	Results	128
A.2	SSS Transplantation Evaluation	130
A.2.1	Evaluation Design	130
A.2.2	Evaluation Results	131
A.3	Cross-lingual Emotion Transplantation	132
A.3.1	Experimental Evaluation	132
A.3.2	Perceptual Evaluation and Covariance Analysis	137
A.4	Transplantation System Performance Analysis	138
A.4.1	Evaluating the distance between parallel contexts	138
A.5	MRHMM-based transplantation	139
A.5.1	Perceptual Evaluation	141
A.5.2	Results	142
B	Software and Versions	146
B.1	Speech Synthesis	146
B.2	Genre Prediction	146

List of Figures

2.1	Sample speaking styles categorization, from [Eskénazi, 1993].	10
2.2	Two implementations of Unit Selection concatenation algorithm. The baseline (left), and a decision-tree clustering based one (right), from [Zen et al., 2009] . .	13
2.3	Generic HMM-based Speech Synthesis diagram, extracted from [Zen et al., 2009]	14
2.4	Schematic of the feature synthesis process, from HTS Documentation.	15
2.5	Schematic of DNN-based TTS, from [Zen et al., 2013].	18
2.6	Speaker interpolation schematic at training time (left) synthesis time (right), adapted from [Yoshimura et al., 1997].	20
2.7	Schematic of the CSMAPLR adaptation algorithm, from [Yamagishi et al., 2009].	22
2.8	Text pre-processing schematic.	24
2.9	Schematic of the SVD decomposition of the WTDM, from [Echeverry-Correa, 2015].	29
2.10	Schematic of the truncated SVD, from [Echeverry-Correa, 2015].	30
3.1	Proposed system for paralinguistic features control through adaptation.	45
3.2	Results of the speech quality evaluation for natural voice in MOS scale.	50
3.3	Results of the speech quality evaluation for the adaptation-based control system synthetic speech in MOS scale.	50
3.4	Results of the speech quality evaluation for interpolation-based control system synthetic speech in MOS scale.	51
3.5	Plots of the perceived emotional strength of the four emotions in the paralinguistic features control system evaluation in a 1 to 100 scale.	52
4.1	Schematic of the emotion transplantation method.	57
4.2	Step by step block diagram of the emotion transplantation method. The ellipses represent the speaker models and the arrows the adaptation transforms.	59

4.3	Boxplots for the results of both speech quality and emotional strength for the first transplantation evaluation. The blue bars represent the average emotion transplantation system and the red bars represent the neutral system.	65
4.4	Boxplots for the results of both speech quality and emotional strength for the second transplantation evaluation. The green bars represent the proposed transplantation system and the red bars represent the neutral system.	66
4.5	Boxplots for the results of both speech quality and emotional strength for the final transplantation evaluation. The green bars represent the proposed transplantation system and the blue bars represent the average emotion transplantation system. .	68
4.6	Picture of the robotic assistant, Urbano.	70
4.7	Expressive faces for the robot in the following order: neutral, happiness, sadness, anger and surprise.	70
5.1	F0 vs. rhythm map of the 4 recorded speaking styles in SSS.	76
5.2	Schematic of the proposed discrete approach to the expressive TTS system. . . .	79
5.3	Schematic of the proposed semi-continuous approach to the expressive TTS system.	82
5.4	Schematic of the proposed semi-continuous approach to the expressive TTS system.	84
5.5	Results in MOS scale of the adequacy evaluation for the discrete system.	87
5.6	Results in MOS scale of the expressive intensity evaluation for the discrete system.	87
5.7	Results in MOS scale of the quality evaluation for the discrete system.	88
5.8	Results in MOS scale of the adequacy evaluation for the semi-continuous system.	89
5.9	Results in MOS scale of the expressive intensity evaluation for the semi-continuous system.	89
5.10	Results in MOS scale of the quality evaluation for the semi-continuous system. .	90
5.11	Schematic of the proposed continuous approach to the expressive TTS system. . .	91
5.12	Results in MOS scale of the adequacy evaluation for the continuous system. . . .	94
5.13	Results in MOS scale of the expressive intensity evaluation for the continuous system.	94
5.14	Results in MOS scale of the quality evaluation for the continuous system.	95
A.1	Results of the perceptual test comparing the user preferences between transplantation and copy of prosody for all test speakers and styles on C-ORAL-Rom database.	127

A.2	Step by step block diagram of the emotion transplantation method. The spheres represent the speaker models and the thin arrows the adaptation transforms. The state mapping process is represented by the rectangle and the result of the mapping as the thick arrow. The two colors also represent the different languages. . .	133
A.3	Illustrations of the three different considered cross-lingual experimental variations. Black filled surfaces represent nonzero values.	133
A.4	PCA projection of the different adaptation threshold experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.	135
A.5	PCA projection of the different adaptation block size experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.	136
A.6	PCA projection of the different adaptation bandwidth experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system, constant across experiments.	137
A.7	Flowchart of the proposed MRHMM-based emotion transplantation system. . .	140
A.8	Similarity results in percentage of the first perceptual MR-HMM based transplantation evaluation with the transplanted emotion control factor set at 0.75. . .	144
A.9	Similarity results in percentage of the second perceptual MR-HMM based transplantation evaluation with the transplanted emotion control factor set at 1.0. . .	145

List of Tables

2.1	Example of a 5 by 5 Latin square	38
2.2	Mean Opinion Score value association.	39
3.1	Systems evaluated in the paralinguistic features control system.	47
3.2	Confusion matrix for the emotion identification rates for natural voice in the paralinguistic features control system evaluation in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.	48
3.3	Confusion matrix for the emotion identification rates of the adaptation-based system voice positive emotions in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.	48
3.4	Confusion matrix for the emotion identification rates of the adaptation-based system voice negative emotions in percentage. Rows show the chosen emotion and columns the actual emotion.	48
3.5	Confusion matrix for the emotion identification rates of the interpolation-based system voice positive emotions in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.	49
3.6	Confusion matrix for the emotion identification rates of the interpolation-based system voice negative emotions in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.	49
3.7	Recognition rates in absolute percentage values averaged across all emotions. . .	49
3.8	Speech Quality results in MOS scale averaged across all emotions.	51
3.9	Pearson correlation coefficient between the perceived emotional strength results and the theoretical perfectly linear system.	52
3.10	Results relative to natural speech for the adaptation-based expressive control system and traditional speech synthesis[Iberspeech2012, 2012]. Bold values represent interesting values.	53

3.11	Mean results in MOS scale comparing a traditional, neutral TTS system (BASE-HIFI) and an emotional system using the proposed method (NEMOHIFI) extracted from [Lutfi et al., 2013]. X denotes not significant differences.	54
4.1	Summary of the results for the preliminary evaluation [Lorenzo-Trueba et al., 2013a] on emotion transplantation for speech quality, emotion identification, emotional strength and similarity to the target speaker for a transplantation control ratio of 1.00. Results are relative to those of the natural voice and showed in percentage values.	61
4.2	Sample emotional utterances used for the proposed transplantation system preference evaluations translated from Spanish	63
4.3	Sample emotional utterances used for the proposed transplantation system preference evaluations in Spanish	63
4.4	Source and target data used for training the adaptation functions for the 3 systems in the paralinguistic features control system evaluation.	64
4.5	Complete characterization of the speakers used in the emotion transplantation evaluations. In the emotions section, A stands for anger, H for happiness, S for sadness, U for surprise and N for neutral.	64
4.6	Results for the first transplantation evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant. Quality and strength are represented in MOS scale, preference in percentage. . .	65
4.7	Results for the second transplantation evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant. Quality and strength are represented in MOS scale, preference in percentage. . .	66
4.8	Confusion matrix in percentages for the multiple choice preference test considering all four transplanted emotions and neutral voice. Rows show the chosen preferred system and columns the actual transplanted emotion.	67
4.9	Results for the final transplantation evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant. Quality and strength are represented in MOS scale, preference in percentage. . .	69
4.10	Sample happiness dialog used in the robot evaluation, translated from Spanish. The italics text shows the emotional turns in the dialog.	71
4.11	Sample happiness dialog used in the robot evaluation in Spanish. The italics text shows the emotional turns in the dialog.	71

4.12	Results in percentage for the first preference evaluation with the robotic assistant. The X represents results that are not statistically significant.	72
4.13	Results in percentage for the second preference evaluation with the robotic assistant. The X represents results that are not statistically significant.	72
5.1	Detailed description of the SSS database.	77
5.2	Sample test sentences from SSS. News and political speech samples are shortened for representation purposes.	77
5.3	Train texts genre predictor confusion matrix in percentage. Speech stands for political speech and sports for live sports broadcasts. Rows show the predicted style and columns the actual style.	85
5.4	Evaluation texts genre predictor confusion matrix in percentage. Speech stands for political speech and sports for live sports broadcasts. Rows show the predicted style and columns the actual style.	85
5.5	Results in MOS scale for the perceptual evaluation of the SSS database with error margins for a 95% confidence rate. Intensity stands for expressive intensity and quality for speech quality. Also, speech stands for political speech and sports for live sports broadcasts.	86
5.6	Percentage of synthesizable test sentences for each system-style pair. C_Normalized represents the normalized implementation of the continuous system and Continuous the non-normalized one.	93
5.7	Adequacy results in MOS scale for all the compared expressive TTS systems. . .	95
5.8	Perceived expressive intensity results in MOS scale for all the compared expressive TTS systems.	96
5.9	Speech quality results in MOS scale for all the compared TTS systems.	96
A.1	Description of the training data for the speaking styles transplantation evaluation. All speakers are male and speaking styles are extracted from the Spanish section of C-ORAL-Rom.	128
A.2	Preference results in for the evaluated utterances in percentages in terms of expressiveness and similarity. The results show the preference for the transplanted system over the copy of prosody on C-ORAL-Rom database.	129
A.3	Results in adequacy in MOS scale for the different systems averaged between the 3 target speakers in the first speaking styles transplantation evaluation on C-ORAL-Rom database.	129

A.4	Results in similarity in MOS scale for the different systems averaged between the 3 target speakers in the first speaking styles transplantation evaluation on C-ORAL-Rom database.	129
A.5	Results in preference in percentage towards the proposed system for the adequacy and similarity in the SSS transplantation evaluation. The X represents a not statistically significant result.	131
A.6	Results in MOS scale for speech quality for both the proposed and neutral system in the SSS transplantation evaluation. A X represents a not statistically significant result.	131
A.7	Results in MOS scale for perceived expressive strength for both the proposed and neutral system in the SSS transplantation evaluation. A X represents a not statistically significant result.	132
A.8	Description of the training data for the cross-lingual emotion transplantation. All speakers are male.	134
A.9	Covariance analysis results for the three cross-lingual experimental environments.	138
A.10	Percentage of contexts, with the error margin for a 95% confidence, that presented errors by feature stream and context distance.	139
A.11	Speech data used to train each background emotion model in the MR-HMM based transplantation.	141
A.12	Confusion matrix in percentage of the emotion identification rates for the source speaker (ROS) in percentage for the MR-HMM based transplantation. A stands for anger, H for happiness, N for neutral, S for sadness, U for surprise and O for other. Results below the random threshold are omitted.	142
A.13	Confusion matrix in percentage of the emotion identification rates for the target speaker (UPC) in percentage for the MR-HMM based transplantation. Column headers are analogous to table A.12.	143

Chapter 1

Introduction

1.1 Motivation

Since the advent of the Information Age, computers have begun playing constantly increasing roles in the daily life of humans in the developed and developing countries: internet, smart phones, tablets, virtual agents and even public transport ticket vending machines are just a few examples of machines we have included in our daily routine. In consequence, numerous fields of study around these machines have appeared. Among them, the study of human-machine interactions and interfaces pose a definite challenge, as providing simple and efficient communication interfaces could significantly reduce the gap in technology usability between those trained in it and those untrained.

These communication interfaces, known as Human-Machine Interfaces (HMI), have been studied under different approaches. Some approaches try to consider the machine as a part of the human itself, giving birth to the idea of brain-computer interfaces (BCI) [Wander and Rao, 2014; Ortiz-Rosario and Adeli, 2013], combining medicine and engineering with applications such as biomechatronic hands [Mattar, 2013] that can be controlled with just with thoughts, or artificial eyes [Luo and da Cruz, 2014] and ears [Käthner et al., 2013] to give sight or hearing back to people who lost it.

Other studies, on the contrary, approach the communication process as if it were a typical inter-human communication. Applications such as virtual agents or avatars [Yaghoubzadeh et al., 2013; Krum et al., 2014; San-Segundo et al., 2012], dialog systems [Lutfi et al., 2013] are clear

examples of how HMI can be used as a tool for improving our quality of life. This later approach intention is to normalize the interaction with computers to the point that there is no difference between communicating with another human and communicating with a computer. Among the possibilities, using speech as a means of giving and receiving information comes naturally and is commonly referred to as Speech Technology, and is a field of study with many different sides: speech recognition [Woodland et al., 1995; Graves et al., 2013], speech synthesis [Tokuda et al., 2002; Watts, 2012], speaker and language recognition [Campbell Jr, 1997; D'haro Enríquez et al., 2012], and so on.

When talking about speech synthesis the objective is to produce artificial human speech by means of either hardware or software technologies. A broader application enclosing speech synthesis is the text-to-speech synthesis (TTS), which aims to translate natural language texts into said synthetic speech. This problem has been pursued for many years beginning with the first hardware vocal tract models in 1779 by Christian Kratzenstein, and systems nowadays can produce synthetic voices with high enough quality that the academia's focus has shifted into new problems: expressive speech synthesis [Schröder, 2009; Iriondo et al., 2007], expressiveness transplantation [Lorenzo-Trueba et al., 2013a,b], voice conversion [Erro et al., 2013b; Stylianou et al., 1998], and so on. The main reason for this shift in objectives is that even if the voice is high quality enough, TTS systems are not capable of producing natural, real-life sounding speech, which automatically hinders their implementation possibilities.

Adding expressive capabilities to speech technology greatly expands the potential fields of use: beginning with lie detectors [Gadallah et al., 1999] and emotion detectors [Koolagudi and Rao, 2012], continuing with realistic expressive dialog systems that adapt to the users emotional state [Bickmore et al., 2012] or systems capable of using emotions for better conveying their actual state [Lutfi et al., 2013] finishing by more commercial applications such as automatic voice castings and synthesis for movies or video games [Obin et al., 2014; Norris et al., 2014] or automatic synthesis of audiobooks with multiple characters [Chen et al., 2013a].

The work proposed in this thesis is enclosed mainly under the field of expressive speech synthesis, and aims to fix one of its main shortcomings: scalability. Human communication is so rich and so deep that it is impossible to imagine obtaining data for every combination of speaker and expressiveness, and that is why we want to obtain a method capable of learning the paralinguistic information of expressive speech, and control it and even transplant it to different speakers for whom we do not have any expressive information.

1.2 Organization

The present PhD thesis is organized as follows. First, chapter 2 talks about the theoretical background of the topics treated in this thesis, chapter 3 describes the work carried out during the development, design and evaluation of the paralinguistic features control system. Chapter 4 and annex A explain the details of the design of the paralinguistic information transplantation system: chapter 4 describes the system and the evaluations carried out to validate the system. Then, in chapter 5 the development of the complete expressive TTS system is explained. Finally, chapter 6 shows the conclusions that can be extracted during the research process presented in this work. Annex A introduces additional work that arose from the transplantation system itself (e.g. cross-lingual transplantation, speaking styles transplantation...).

1.3 Objectives

This thesis aims to fill a relevant hole in the field of expressive speech synthesis: the scalability problem. Human speech is extremely rich not only in its acoustic information but also in its linguistic, paralinguistic and extra-linguistic properties. Speaking rhythm, accents, intonation, etcetera, all contribute to the "color" of a voice, and can carry important information for conveying the desired message. As such, it is natural to think that it is not feasible to record data for every single expressive nuance of speech and for every speaker whose voice we want to feature in our TTS system, the amount of required data would be too high.

We propose the design of a method capable of learning these paralinguistic properties of speech and controlling them, allowing us to control the strength in which they are present in the synthesized speech, enhancing or lessening the expressiveness of the produced voice.

This technique will also allow us to transplant the paralinguistic information into a different target speaker, effectively allowing us to produce expressive voices for any target speaker even if only neutral speech was recorded.

The final objective in this PhD thesis is to provide a complete system, capable of automatically predicting the genre of a text, and by means of natural language processing techniques establish a relationship between text genre and speaking expressiveness, providing a text to speech system that automatically adapts to the task for any desired target speaker, with minimal interaction with the end user.

1.3.1 Design of a New Method for Controlling the Strength of Paralinguistic Properties of Speech

The first objective of this thesis is to design a new method capable of controlling the paralinguistic properties of a collection of utterances. Natural speech utterances are used to train the speech models, whose distributions we are able to transform by means of adaptation techniques. This allows us to manipulate their paralinguistic features, enabling us to modify the expressive strength of the produced voices. For this objective we want to focus on a particular expressiveness: emotions.

The proposed method will be able to:

- Avoid the instabilities that commonly appear when using traditional interpolation approaches, allowing us to maintain speech quality and speaker identifiability while modifying the paralinguistic nuances to a much bigger extent [Lorenzo-Trueba, 2012].
- Compensate the inherent smoothing caused by the Gaussian modeling of HMM-based synthesis by enhancing the expressive strength of the speaker models [Lorenzo-Trueba et al., 2012b].
- Allow the possibility of generating the desired controlled expressive synthetic speech models offline. This means that using expressive voices synthesized by the proposed technique on dynamic and real applications is possible, as most of the processing time of generating the voices can be handled beforehand, instead of controlling it at synthesis time.
- Control expressive strength, meaning that we can give our text to speech systems a higher versatility that mimics more closely human speech. In real life humans do not talk constantly at full-fledged expressiveness, so being able to imitate that behavior is necessary if we want the listener to perceive the system as more natural [Lutfi, 2013].

1.3.2 Design of a New Method for Transplanting Paralinguistic Information between Speakers

The second objective is to devise a new method able to learn the paralinguistic information that we were able to control in the previous method, and then capable of transplanting it to any new desired target speaker. This method should be capable of imbuing the paralinguistic information

into the target speaker even if only neutral data is available by learning it from a source model. Concretely, we will consider two aspects of paralinguistic information, we will study mostly emotions but we also will consider speaking styles and their impact in the acoustic features.

The main aims of the second objective are:

- Learn the paralinguistic information of a prototypical speech synthesis model and by means of chained adaptations, transplant said information into a new target speaker from whom only neutral, read speech data is available [Lorenzo-Trueba et al., 2013b,a]. This method, if successful, would greatly help to alleviate the scalability problems present in nowadays expressive text-to-speech systems, as we would be able to produce expressive speech for any target speaker as long as we have the reference expressive model and a few minutes of neutral speech for the target speaker [Lorenzo-Trueba et al., 2015].
- Control the relevance of the paralinguistic information of both the expressive and the target speaker identity models akin to the expressive strength control proposed in the first objective. This would allow us to produce expressive speech models with the desired strength while also allowing us to prevent the dissipation of the speaker identity by means of the transplantation factor.
- If the baseline transplantation technique provides good results, we also intend to use cross-lingual adaptation techniques to design a cross-lingual expressiveness transplantation method, capable of learning the paralinguistic information of a target expressiveness in a given language, and then transplanting it into a target speaker of a different language [Lorenzo-Trueba et al., 2014].

1.3.3 Design of an Expressive TTS System: From Text Genre to Speaking Styles

The third and final main objective of the present PhD thesis is the design of a complete expressive text-to-speech system. The complete system is the final part of the puzzle of our proposed expressive speech synthesis methodologies, and as it is an application we consider it is most useful when applied to speaking styles synthesis. The system will have to be capable of detecting the genre of the text to be synthesized, and by means of different techniques that prediction would be assigned to an adequate expressive speaking style model, which will be transplanted into the

speaker selected by the user as the target. On this objective we will focus mainly on speaking styles as the desired expressiveness.

The complete system should be capable of:

- Predicting the genre of the text to be synthesized. This can be done either by using traditional information retrieval techniques and likelihood thresholding to give a discrete prediction, or by means of more complex Latent Semantic Analysis (LSA) and vector space modeling to obtain distances between the target text and prototypical genres, which would in turn give us a semi-continuous genre classification.
- Selecting the expressive speech model according to the prediction. Once the genre prediction is available, we intend to assign a correspondence between genre and speaking style speech model, which can also be picked directly from the discrete prediction, or controlled with the distance value of the semi-continuous prediction.
- Synthesizing the input text with the chosen expressiveness and the desired target speaker. The final step is to apply the transplantation methodology explained in the previous objective between the automatically selected speaking style and the chosen target speaker, the output of the system is the synthesized text according to all the requirements with minimal effort from the end-user as there is no need for recording additional expressive data from the target speakers.

A different approach to the discrete and semi-continuous is to use the output of the LSA and Single Value Decomposition (SVD) as the HMM training process features, removing the need of predicting genre as the text would be used directly as the speaking style feature in a completely continuous modeling fashion.

Chapter 2

State of the Art

2.1 Expressive Speech

Human spoken communication is capable of conveying more information than that which is transmitted by the syntactic combination of lexicals. This means that there are more faces to speech than just the words: emotions, speaking styles, accents, etc... Recent studies in expressive speech and language [Schuller et al., 2013; Schweinberger et al., 2014] classify speech into three main aspects: linguistic, paralinguistic and extra-linguistic. Linguistic information is the meaning carried by the words themselves, while paralinguistic communication is associated to dynamic variations in the way of speaking (i.e. emotional state, speaking style...) and extralinguistic information is carried by the longer-term characteristics of speech (i.e. speaker identity, age...). And it is safe to consider that the correct combination between them is what conveys the full meaning of human communication. Consequently, according to the previous definitions, we can consider expressive speech to be a combination of all three of them, but this study will focus on linguistic and paralinguistic information.

2.1.1 Emotions

Definition of Emotions

Emotions are an integral part of human life, and have been a target of study since time immemorial by psychologists and philosophers. A definition of emotion is "mental state that,

consciously or not, can catalyze physiological reactions for appraisal or adaptation purposes" [Douglas-Cowie et al., 2003]. These reactions can be commonly perceived externally and as such are a very relevant part of human communication, reflected for example in changes in how we speak, resulting in what we define as emotional speech.

Defining what is an emotion, then, is not the main problem, but detecting how they appear is. Humans, by nature, have social barriers that prevent full-blown emotions from appearing in daily life [Douglas-Cowie et al., 2003], as they are perceived as syndromes or hints of a underlying psychopathology, so they are not an ethical target of research. Fortunately, if we consider the psycho-evolutionary theories of emotional research [Ekman, 1999], "emotions are directly linked to systems that evolved to deal with biologically significant situations; and they are therefore universal". In that sense, we can consider that the physical expression of emotions: the supra-segmental features of speech, facial expression, bodily language... can be universally understood and as such can be mimicked to convey the desired emotion. This characteristic of emotions proves extremely useful as it allows us to study them in a controlled environment without losing information, and can be copied into virtual agents (robotic faces, emotional synthetic speech, and so on) to produce synthetic emotions.

Characterization of Emotions

Psychologists through the ages have proposed many theories on how to classify emotions [Plutchik, 1980; Parrott, 2001; Ekman et al., 1972], and nowadays most of them agree on a clear separation between primary (basic) and secondary (derived) emotions. Primary emotions are considered to have been developed evolutionary as responses to particular aspects of the environment, and have been derived from the observation of facial expressions [Ekman et al., 1972] and could be considered basic for human nature, therefore universal: anger, disgust, fear, happiness, sadness and surprise. Secondary emotions are, consequently, the result of the combination between the primary emotions, and it is mainly in the mechanism of these combinations and in how secondary emotions can be classified and produced that the theories tend to differ.

A traditional approach to emotion modeling [Wundt, 1897] proposed a three-dimensional abstract space in which to place them: "pleasurable versus unpleasurable", "arousing or subduing" and "strain or relaxation", nowadays referred to respectively as "Valence", "Arousal" and "dominance" [Mehrabian, 1980]. Emotions can be then classified as discrete entities in this space, or they can be projected into a continuous feature space measured from the expression itself.

Emotions in Speech Processing

Recently there has been a surge in artificial intelligence (AI) research, and within it, a deep interest in human behavior modeling and mimicking [Planet and Iriondo, 2012, 2013]. Speech, being one of the more natural way of carrying emotional information in communication, is one of the easier and more important targets of emotional modeling research. Fortunately, the great advance in speech technologies of the last decades has placed speech synthesis technologies in a situation where obtaining good intelligibility levels is reasonably accessible, but increasing synthetic speech naturalness is, and here is with the correlation between emotions and speech become relevant.

Naturalness in speech synthesis systems clearly relates with how spontaneity and speech variety are handled in the interaction process. Effects such as non-standard pausing, hesitation, co-articulation, pitch variability, intonation... all affect the expressiveness and naturalness in speech and a correct modeling and production greatly increases the overall quality of the system [Monzo et al., 2014].

The interest in emotions in this thesis lies in their universality and controllability, making them exceptionally good targets for expressive text-to-speech research.

2.1.2 Speaking Styles

Speaking Styles Definition

Speaking style is a very broad term used to define how the "way of speaking" shifts according to both linguistic environment and social context constrains variations [Argente, 1992], and can be associated to the difference between "language" and "speech" given in early speech studies [Sapir, 1921]. Language is considered to be patterned, constrained by rules, while speech is the unpatterned, unpredictable and apparently ungoverned manifestation of language. In conclusion, we can consider that a speaking style is any particular manifestation of speech associated to a particular situation that will modify the way of speaking compared to what formal language stipulates.

Characterizing Speaking Styles

Speaking style, in opposition to emotion, has a much more diffuse characterization because of its already vague definition, and is typically done by contrast of styles instead of using an absolute mapping [Abe, 1997]. For example we could consider speaking styles in the sense of spontaneous vs. scripted styles: interviews, sports commentaries vs. news casting, or we could consider the mode of discourse: monologue vs. conversation, and so on [Llisterri, 1992]. To sum up, speaking styles serve as a way of comparing manners of speaking between each other in order to provide separable categories [Eskénazi, 1993] such as the one that can be seen in figure 2.1.

Figure 2.1: Sample speaking styles categorization, from [Eskénazi, 1993].

Attempts to characterize formally the whole spectrum of speaking styles in segmental and supra-segmental features (i.e. pitch, phoneme durations, speaking rate... as examples of supra-segmental features, and phoneme elision, deletion, co-articulation... as examples of segmental features) to better define them for the TTS problem has shown that speaking styles tend to be better captured at the supra-segmental level, especially by fundamental frequency, duration, speaking rate and pauses. Nevertheless, information of vowel reduction or elision also seem to provide useful information [Llisterri, 1992].

Speaking Styles in Speech Processing

Speaking styles, once again because of its huge scope, is a significant problem for speech processing as it relies heavily in the three aspects of communication previously mentioned: linguis-

tic, paralinguistic and extra-linguistic. Then, because it is not viable to focus on speaking styles as a whole, studies are typically focused on particular characteristics of one of the aspects: for example prosodic mapping of the paralinguistic features, speaker identification making use of extra-linguistic information or genre classification as a linguistic approach.

In the end, speaking styles is something that must be paid attention to regardless of the application as long as it has something to do with speech [Veiga et al., 2012; Moinet et al., 2013; Obin, 2011; Picart et al., 2013]. As this PhD thesis aims to provide improvements to text-to-speech synthesis, and more concretely to their expressive capabilities, it is natural that they also become a very significant target of our study.

2.2 Text to Speech Synthesis

2.2.1 Introduction

Speech synthesis is a field of speech processing that has been receiving attention for a very long time. Since the first archaic implementations of hardware speech synthesizers by Christian Kratzenstein with his model of the human vocal tract in 1779, and Wolfgang von Kempelen with his "acoustic-mechanical speech machine" in 1791, speech synthesis has grown substantially to the point nowadays that speech quality is not the main challenge. In more recent times, concatenation of natural speech units [Olive, 1980; O'Shaughnessy et al., 1988] became the dominant technique after rules-based formant synthesis [Klatt, 1980] reached their peak.

Concatenation of natural speech, albeit potent, was too rigid because they depended usually on diphones or syllables, which in the end proved not to be good enough. That is why, together with the increases in computational power, there was a progressive increase in the amount of units in the databases, and nowadays concatenation synthesis systems have tens of thousands of triphone units. This large database concatenation systems are now called Unit Selection (US) systems [Hunt and Black, 1996]. But, as US systems rely on natural voice for speech production, their quality is very high but their adaptability is extremely low, that is why explicit modeling techniques such as Statistical Parametric Speech Synthesis (SPSS) [Black et al., 2007] were born.

Nowadays both US and SPSS coexist, each with their advantages and disadvantages, with hybrid techniques [Sun et al., 2012; Tiomkin et al., 2011; Sainz et al., 2011] trying to bridge the gap between them, and with processing techniques such as voice conversion [Erro et al., 2013a;

Helander et al., 2012] that are sometimes used to alter the voices to provide different speaker identities or expressiveness.

Another technique that has had a resurgence starting a few years ago is the neural network based speech synthesis by means of deep neural networks (DNN) [Zen et al., 2013; Wu et al., 2015] or recurrent neural networks (RNN) [Fan et al., 2014; Zen and Sak, 2015]. This resurgence has occurred mainly thanks to the appearance of efficient unsupervised learning algorithms (deep belief networks, DBN [Hinton et al., 2006]) that revolutionized the approach to initializing the weights of multilayer neural networks, thus solving the optimization problems that multilayer perceptrons (MLP) had in the past.

2.2.2 Unit Selection Based TTS

Unit selection text-to-speech [Campillo and Banga, 2011] is based on the concatenation of natural speech segments by carefully selecting them from a pre-recorded database in order to provide very natural-sounding synthetic speech. The process by which the units are selected from the database is in charge of selecting a segment with the best [Clark et al., 2007; Narendra and Sreenivasa Rao, 2013]: context sequence, accent placement, pitch contour, phone durations, etc.. Evidently, the problem is extremely complex and grows more and more complex with the size of the units database.

With the surge in use of dynamic programming and the increase in computation powers, the database searching algorithm was improved and conformed into a two-fold cost: the target cost and the concatenation cost. Target cost measures how good is the selected unit match to the desired one and is typically based on linguistic features, while concatenation cost is calculated on the goodness of the union between the selected units and its adjacent ones, typically considering the acoustic features. The unit selection process is then in charge of optimizing both costs at the same time, the result being the best unit match. Two different implementations of this algorithm can be seen in figure 2.2, the left one [Hunt and Black, 1996] searching through the whole units database and the right one [Black and Taylor, 1997] making use of decision trees in order to reduce the complexity of the search.

Using natural speech at its core is at the same time the best feature and the biggest shortcoming of US systems: while it allows us to obtain extremely high quality synthetic speech, if we want to produce something that is too far from what is available in the database we have to start manipulating the waveforms and with it the risk of introducing artifacts rises sharply. Certainly

Figure 2.2: Two implementations of Unit Selection concatenation algorithm. The baseline (left), and a decision-tree clustering based one (right), from [Zen et al., 2009]

there are many techniques (PSOLA modifications [Mousa, 2010], sinusoidal modeling [Macon and Clements, 1996; Türk and Schröder, 2010]) capable of modifying the pitch and duration of the available units in order to obtain something closer to what is needed, but they are not enough to compensate the shortcomings of not having a bigger database, which is why recently SPSS systems and hybrid US-SPSS systems have become much more predominant.

2.2.3 Statistical Parametric Speech Synthesis

Statistical Parametric Speech Synthesis (SPSS) is completely different to US in the sense that while in the former speech is just replicated from a database, SPSS aims to generate speech purely from statistical models of the waveforms, only relying on natural speech for training those models. Typically, Gaussian models are obtained from the training data by means of the Expectation Maximization (EM) process [Moon, 1996] and the desired speech signal is estimated through a Maximum Likelihood (ML) algorithm [Dempster et al., 1977]. And, although different generative models exist, Hidden Markov Models (HMM) are by far the most widely known and used.

HMM-Based Synthesis

HMM-based speech synthesis [Takaki et al., 2013] is the most known and widely used parametric speech synthesis system, and we can see the typical architecture in figure 2.3. From the parameter extraction, we can see how both excitation (typically F0 or its logarithm) and spectral (Line Spectral Pairs (LSP) [Kim et al., 2006; Jang and Cho, 2011], Mel Cepstral Coefficients (MCEP) [Fukada et al., 1992; Valentini-Botinhao et al., 2014]) features are extracted and modeled together with alignment information from the database into a set of HMM. State of the art excitation models such as STRAIGHT [Kawahara et al., 2001], which will be used for the purposes of this PhD thesis, also consider the use of a mixed excitation model weighting a combination of F0 and aperiodicity bands, together with the dynamic features of speech (first and second order derivatives of the extracted parameters).

Figure 2.3: Generic HMM-based Speech Synthesis diagram, extracted from [Zen et al., 2009]

The HMM training process always relies on very rich contexts (typically pentaphonemes) and syllable, word, phrase and utterance level contextual information [Okuda et al., 2002], which are then clustered into context-dependent multivariate Gaussian distributions separated by means of

decision trees obtained by the Minimum Description Length (MLD) criterion [Hu et al., 2014] to cluster the contextual information. Also, in order to be able to provide richer information for the voiced F0 frames, they are modeled using Multi-Space probability Distributions (MSD) [Shen et al., 2011].

Up until now the differences with traditional speech recognition HMM-based systems are minimal: slightly different features, richer contexts... but the biggest difference lies in durations modeling. In traditional HMM-based modeling, durations are modeled just by the transition probabilities of the HMM states. For speech synthesis, it was proven that using explicit distributions for durations too in what is called Hidden Semi-Markov Models (HSMM) allows for much greater accuracy in speech production. This is so because while a traditional transition matrix optimal value is always one, if we use an explicit modeling such as a Gaussian distribution, the optimal value can be adapted accordingly to the distribution, greatly increasing its versatility. Additionally, it allows the system to estimate simultaneously the state output and state duration [Zen et al., 2007].

Figure 2.4: Schematic of the feature synthesis process, from HTS Documentation.

The synthesis part of the process (figure 2.4) consists in finding the optimal state outputs for the given desired features and contexts. This is done in two steps: first the text to be synthesized is converted into a context labels sequence with all the needed information as used in the training process. Then, a speech parameter generation algorithm [Tokuda et al., 2000] is used to provide

the input to the HMM models. Then, by means of a ML estimation process, the optimal state outputs are estimated and maximized. It is important to note that because not only the features but also the first and second order derivatives are obtained, the final features trajectory will have much smoother transitions, providing more stable outputs. Finally, the desired speech waveform is synthesized from the generated spectral and excitation parameters by means of excitation generation and filtering processes such as Straight or Mel Log Spectrum Approximation (MLSA) [Zen et al., 2011; Imai et al., 1983].

2.2.4 Strengths and Weaknesses of the Techniques

The advantages and disadvantages of both US and HMM-based systems have already been mentioned throughout the explanation, and they happen to be quite complementary:

Unit selection systems are capable of providing extremely high quality speech, but at the cost of a very rigid structure that has very little adaptation capabilities. Also, unit selection suffers from irregular quality, presenting big variation between contexts available in the corpus and those unavailable. Both disadvantages together mean that the US systems main disadvantage lies in their lack of scalability.

HMM-based systems due to their parametric nature are very easily adaptable by means of the abundant adaptation techniques that will be introduced in section 2.3.2. Also, the computational footprint is small once the HSMM have been trained because the generation process is much simpler, together with the fact that HMM-based synthesis have been used for speech recognition tasks for a really long time, so they are very optimized. Unfortunately, the shortcoming is that the speech quality and naturalness attainable is smaller than that attainable with US systems.

In the end, we can see how depending on the requirements of the intended application, one would pick US or HMM-based systems. Nonetheless, there are studies on how to reach the best of both worlds: hybrid systems attempt to mix both techniques to lessen the effect of their shortcomings, and voice conversion techniques study how the speech signal can be morphed without having to rely on adaptation techniques.

Hybrid Systems

There are many variations of hybrid systems. The most common hybrid systems substitute the target selection of the unit selection process by HMMs, which are used to generate the acoustic features that are used for the target selection cost calculation [Chen et al., 2013b]. Other systems attempt more extreme variations such as mixing natural speech segments with HMM generated segments to fill the missing gaps [Okubo et al., 2006], or propose the use of HMM models to smooth the natural speech units [Plumpe et al., 1998].

These systems are able to alleviate some of the irregularity problems intrinsic to US systems, and are also able to provide even higher quality speech, winning many speech synthesis competitions through the years [King and Karaiskos, 2013]. Even so, the scalability problem is not really solved as the reliance in natural speech segments is still there, added to its inherent lack of adaptability.

Voice Conversion

Voice conversion techniques were developed as a means of morphing the voice from a speaker to make it sound like another speaker (i.e. Speaker Conversion [Stylianou et al., 1998; Kain and Macon, 1998]. As it is commonly accepted that speaker identity is mainly captured by the spectral features, most of such techniques are focused on spectral morphing [Desai et al., 2010; Erro et al., 2013b], frequency warping [Erro et al., 2010a] or vocal tract manipulations [Erro et al., 2012]. Regardless, it has been seen that voice conversion can be used with more applications in mind than just speaker conversion: it can be used to convert the voice in general.

Voice conversion in general can also be done in prosodic features, and some studies considered fundamental frequency transformations [Wu et al., 2010], duration morphing [Srikanth et al., 2012], or even mixed prosodic and spectral transformations [Aihara et al., 2012]. This gave birth to the idea of producing expressive voices from non-expressive speech by means of voice conversion, and a few of such studies have been published recently [Türk and Schröder, 2010].

2.2.5 Neural Network based synthesis

Conventional SPSS typically relies on decision trees to map from the linguistic contexts extracted from the text to the probability densities of the speech parameter models we use to synthesize.

This approach, while effective, has some limitations. First of all it is inefficient for expressing complex context dependencies, which limit the model's long-term modeling capabilities. Another inherent problem of decision trees is that they partition the input space, which automatically fragments the training data potentially inducing over-fitting [Kotsiantis, 2013]. The main advantage that DNNs provide to solve the aforementioned problems is that, with multiple hidden layers, they are capable of expressing complex functions more efficiently than what a binary tree can do as long as the layers have enough units. Moreover, they are also very efficient at jointly modeling the different streams [LeCun, 2012].

Fig. 1. A speech synthesis framework based on a DNN.

Figure 2.5: Schematic of DNN-based TTS, from [Zen et al., 2013].

In figure 2.5 we can see a common framework for DNN-based TTS. In these systems we extract the linguistic features as with traditional SPSS, which are then fed to the input layer of the DNN. The number of hidden layers and units depends on the models and must be optimized, but a common optimal structure is 4 layers with 512 units per layer for a total of more than 1 million parameters to be estimated [Zen et al., 2013]. In order to estimate that amount of parameters and to avoid the vanishing gradient problem caused by back-propagation [Mikolov et al., 2014], restricted Boltzmann machines (RBM) are used to initialize each individual layer. These systems have been shown to perform significantly better than traditional HMM-based TTS, mainly due to

the better MCEP prediction. Since the basic DNN structures were proposed, many improvements and implementations have been considered such as stacked bottleneck features [Wu et al., 2015], DNN-based sinusoidal speech synthesis [Hu et al., 2015] and even initial attempts at speaker adaptation [Fan et al., 2015].

Unfortunately, even if DNN-based synthesis outperforms HMM-based synthesis, they both suffer from using short units (frames and states respectively) as modeling units. This limits the model’s capabilities to capture co-articulation effects or long-term prosodic trajectories. RNNs, on the other hand, due to their recurrent nature, are theoretically capable of capturing information from anywhere in the feature sequence, solving the short-time effect problem [Graves et al., 2013]. Particularly, the long short-term memory (LSTM) approach has been successfully used to efficiently synthesize long-term dependencies significantly improving naturalness [Fan et al., 2014; Zen and Sak, 2015].

Regarding expressive TTS, neural network based approaches are still very limited. This is mainly because the amounts of data required to train a DNN are significantly higher than those required to train an expressive HMM model, especially if we think about the advantages that adaptation techniques provide in that sense. Nevertheless, we already mentioned that speaker adaptation techniques are being developed and improved everyday for DNN-based approaches [Fan et al., 2015], and some sentence-level speaking styles studies are also beginning to appear [Watts et al., 2015].

2.3 HMM-Based Synthesis Adaptability Techniques

Adaptability in HMM-based systems can take on many different forms: from interpolation techniques capable of modifying state outputs on the fly, to adaptation techniques capable of modifying the speaker models to resemble new observations, and even transplantation techniques capable of changing the models to produce new data according to certain patterns that have not been observed in the intended task.

2.3.1 Interpolation Techniques

Speaker interpolation in HMM-based synthesis has been used successfully as a way of doing speaker conversion [Iwahashi and Sagisaka, 1994; Yoshimura et al., 1997], expressiveness con-

trol [Yamagishi et al., 2007; Hsu and Chen, 2012] and similar approaches. The main feature of speaker interpolation techniques is that it relies in linear combination of the observations of the models to be interpolated. The interpolation therefore can be applied both at training time or at synthesis time.

Figure 2.6: Speaker interpolation schematic at training time (left) synthesis time (right), adapted from [Yoshimura et al., 1997].

As seen in figure 2.6, there are two approaches to traditional interpolation: we could interpolate the features before the training process so we obtain an already interpolated model from "made-up" data, or we could do the interpolation at synthesis time, where we obtain the state outputs for the desired utterance and apply an interpolation between the state outputs of both model. Current Voice Cloning ToolKit (VCTK) implementation¹ uses the second approach, and it is the one we will consider when comparing systems with our proposed system in chapter 3.

The main advantage of using interpolation is that the results attainable are very good for how easy to implement it is. Not only can we produce new speaker models but it can be used successfully as a expressive strength control technique. Sadly there are limitations inherent to modifying state outputs as we will explain in chapter 3 that we will attempt to solve in the thesis PhD proposed in this document as our first objective.

2.3.2 Adaptation Techniques

It is easy to modify HMM-based synthesis models in order to adapt them to the desired task. Examples of situations in which this can be very useful are: improving the speaker model by adding additional training data without requiring a full retraining [Leggetter and Woodland, 1995; Gales,

¹Voice Cloning Toolkit for Festival and HTS - <http://homepages.inf.ed.ac.uk/jyamagis/software/page37>

1998; Gauvain and Lee, 1994; Chesta et al., 1999; Shinoda and Lee, 1997; Siohan et al., 2002; Nakano et al., 2006], adapting speaker dependent data from background Average Voice Models (AVM) [Yamagishi et al., 2003], or even transplanting expressiveness into the speaker models as proposed in this paper.

Generally, when talking about Speaker Adaptation in Speech Synthesis it refers to the process of generating the voice of a desired target speaker from an AVM, and there are different ways to obtain those AVM depending on the intended application. For example, Speaker Adaptive Training (SAT) [Anastasakos et al., 1997] is capable of training average models that excel in mixing speakers with very different characteristics while minimizing the effects of inter-speaker variances through an iterative process, and is capable of providing very high quality voices even if the target voice is very different from the sources [Yamagishi et al., 2009]. Other common techniques are Cluster Adaptive Training (CAT) [Gales, 2000], Eigenvoices [Shichiri et al., 2002] or Multiple-Regression HSMMM [Nose et al., 2007] are very robust against noise as the final model is a combination of the clean source training models, ignoring the noise in the adaptation data and thus requiring less adaptation data than the traditional AVM models. For the proposed task of expressiveness transplantation it becomes very important to both keep the target speaker identity and the source expressiveness, and thus the projective approaches were rejected in favor of a SAT + Speaker Adaptation approach.

For the Speaker Adaptation process from the AVM, there are two main situations to consider: the first is adaptation with enough data to provide robust estimations, typically solved using ML approaches and more commonly maximum likelihood linear regression (MLLR) [Leggetter and Woodland, 1995; Gales, 1998]. Secondly there is the possibility of adapting from sparse data, for which ML is unable to provide robust estimations, originally solved by using prior information in the training process (MAP, maximum a posteriori) [Chesta et al., 1999]. More frameworks have been proposed such as Cepstral mean normalization [Liu et al., 1993], stochastic matching [Sankar and Lee, 1995], vector field smoothing [Tonomura et al., 1995], but they are beyond the purpose of this research.

While MAP is very useful for speaker adaptation in speech synthesis because of its reliability when only having sparse adaptation data, the improvements are typically smaller when compared to MLLR, mainly because in conventional MAP the HMM parameters of the different speech units are assumed to be independent. This means that unless an observation of a particular speech unit is present in the adaptation data, it will not be modified. In order to solve that limitation, a MAP technique that makes use of the structural information of the HMM trees was proposed

[Shinoda and Lee, 1997]. Conversely, MLLR presents the problem of over-training, as it adapts the model as a whole instead of only the observed units. In the end, algorithms that combine the best of both worlds have been proposed, algorithms that exploit the structural information of the trees while using prior information successfully even when the adaptation data is large such as SMAPLR [Siohan et al., 2002].

Focusing on emotional speech adaptation, it has been proved that it is very important to consider not only the means of the HMM Gaussian Distributions but also the variances. This means that it is necessary for the adaptation algorithms to be more complex, or "constrained" as it is called. Ultimately, constrained structural maximum a posteriori linear regression (CSMAPLR) has been proposed and has been proven to be extremely successful for speaker adaptation, particularly when adapting from average voice models [Yamagishi et al., 2009].

Figure 2.7: Schematic of the CSMAPLR adaptation algorithm, from [Yamagishi et al., 2009].

CSMAPLR (figure 2.7) consists in applying the structural MAP criterion (SMAP) [Shinoda and Lee, 1997] to the CMLLR adaptation algorithm [Gales, 1998] and using the recursive MAP criterion [Chesta et al., 1999] to estimate the transforms for simultaneously transforming the mean vectors and covariance matrices of the state output and duration distributions of the speaker model.

There are three main reasons to explain why we will only consider CSMAPLR as the adaptation technique for the rest of the present document. First of all is the aforementioned capability

of not only adapting the mean vectors but also the covariance matrices. The second reason that differentiates CSMAPLR from the more traditional CMLLR adaptation, is that CSMAPLR makes use of the linguistic information of the regression tree by doing recursive MAP-based estimation of the transformation matrices from the root of the context decision tree to the lower nodes, combining the advantages of SMAP and CMLLR. Finally, the fact that CSMAPLR relies on MAP-based estimations means that it is robust when using sparse adaptation data, which is frequently the case in the emotional speech synthesis task.

2.3.3 Multi-Regression HMM adaptation

Multi-regression HMM (MR-HMM) is a particular kind of adaptation in which the model parameters are adapted depending on auxiliary features instead of the acoustic features themselves. Initially this was developed to exploit the correlation between F0 and the spectral features [Fujinaga et al., 2001], which significantly improved isolated word recognition rates in a speech recognition task. Numerically speaking, the multi-regression formula for M auxiliary features is defined as follows:

$$\mu = r_0 + r_1\epsilon_1 + \dots + r_M\epsilon_M \quad (2.1)$$

Where $r_{0...M}$ are the regression coefficients and $r_{\epsilon...M}$ the M auxiliary features. This particularity can be exploited in speech synthesis to model additional information in the speaker models such as speaking styles or emotions [Nose and Kobayashi, 2012; Ling et al., 2013], and it has been used for applications as varied as generating walking motion models [Niwa et al., 2005]. That is, if we use information that characterizes the desired emotion or speaking style that we want to model, we can use it to model multiple expressiveness in a single model as we will do in sections 5.4 and A.5.

2.4 Genre Classification

2.4.1 Introduction

Genre classification is a particularization of a broader topic: Information Retrieval (IR). Information retrieval is the field in charge of representing, sorting and storing information, where information can be any kind of document, web-site, multimedia file [Manning et al., 2008] and so on. In the case we will be studying, genre classification, the objective is to find the topic of a text whether it is structured (web-pages for example) or unstructured.

Information retrieval techniques in general work under the same principle: documents are mathematically represented in a way that allows the input of a query to rank the documents and provide an output. It is the different mathematical representations, query inputs and ranking methods that provide the versatility of the systems, and we will proceed to analyze the common approaches in the unstructured text processing task.

Text pre-processing

As we are dealing with unstructured text, it is important to apply a text pre-processing procedure, seen in figure 2.8, that will provide as an output a normalized vocabulary list.

Figure 2.8: Text pre-processing schematic.

The first module is in charge of a number of tasks: remove punctuation marks, normalize spacings and extra-lexical features in general, and potentially also text normalization (i.e. handling numbers, dates, abbreviations, initials...) [Lopez Ludeña et al., 2012; San Segundo et al., 2012]. This will provide the rest of the system a normalized text without any unwanted character or feature, so the rest of the process can be more general.

The second module is called the "stopwords removal" block, and removes all the too-frequent words that tend to be present in any text [Serrano et al., 2011]: articles, prepositions, pronouns,

and so on. This is done because they do not provide any discrimination capabilities to the information retrieval task, and will reduce the possible vocabulary terms, potentially increasing task accuracy.

The final step is the lemmatization and stemming, reducing the inflectional and derivationally related forms of words to their common base form (e.g. "am, are, is" are reduced to "be"). Both lemmatization and stemming are ways of applying this process, but by stemming we commonly refer to a heuristic process that truncates the end of the words in order to remove the affixes (which is why it is also sometimes called truncating) [Porter, 1980; Paice, 1994]. Lemmatization, on the other hand typically resorts to vocabulary and morphological analysis of words [Le Roux et al., 2012; Hernández-Figueroa et al., 2013]. The output, which is the final output of the pre-processing system, is a list of words without redundant information, vastly reducing the complexity of the following text processing tasks without having lost any semantic information.

2.4.2 Document representation and weighting schemes

Once the documents have been pre-processed and the vocabulary lists have been obtained, it is easy to obtain a vocabulary list and occurrence counts t_i for all the words in the database of documents d_j . This information, when put in matrix form, takes the name of Term-Document Matrix (TDM):

$$TDM = \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & c_{2,3} & \cdots & c_{2,n} \\ c_{3,1} & c_{3,2} & c_{3,3} & \cdots & c_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & c_{m,3} & \cdots & c_{m,n} \end{bmatrix}$$

Where the rows represent the vocabulary list $V = \{t_1, t_2, \dots, t_m\}$, m the size of the vocabulary, the columns the list of documents $D = \{d_1, d_2, \dots, d_n\}$, n the number of documents in the database, and finally $c_{i,j}$ the number of occurrences of term t_i in document d_j . As this matrix is pretty simple, not taking into account for example the length of any particular document or the database itself, whether the term appears in all documents against appearing in only a few [Dumais, 1991]... its problems are numerous and thus its uses are pretty limited, which is why weighted versions of the matrix have been introduced following different weighting schemes.

Most weighting schemes play around two concepts: specificity and exhaustivity [Spark Jones, 1972; Zhang et al., 2011]. Specificity defines how closely a term is related to the topic it represents while exhaustivity is defined around documents, and how many terms are used to describe it, the more terms the more exhaustive. Weighting schemes aim to modify the TDM while keeping the right balance between specificity and exhaustivity, and are applied both locally and globally as follows:

$$w_{i,j} = l_{i,j} * g_i$$

Where $w_{i,j}$ represents every coefficient of the new Weighted TDM (WTDM), $l_{i,j}$ is the local weighting term, and g_i is the global weighting term. There is no universally best weighting strategy [Manning et al., 2008], and they should all be analyzed to optimize the task to be implemented, which will be part of objective 1.3.3 of this thesis.

Local weighting

Local weighting schemes [Manning et al., 2008] target the document exhaustivity, being applied to the TDM terms $c_{i,j}$ independently for each document, and only taking into account the occurrence of term t_i in that particular document d_j :

- **Binary Weighting:** It is the simplest of the weighting schemes, as it consists in calculating a binary mask for existence of term t_i in the whole document d_j , regardless of occurrence count.
- **Term Frequency:** Calculates the relative term frequency (tf) by normalizing the number of occurrences of term t_i in document d_j , normalizing the occurrences count by the number of terms in the document.
- **Logarithmic Term Frequency:** Calculates a logarithmic version of tf in order to reduce the variance in the measures.
- **Augmented and Normalized Term Frequency:** Assigns a minimum value to all terms that appear in the document, adding extra amounts to that value for each additional time the term appears. It must be noted that there is no implicit normalization due to document length, and it must be accounted for by manipulating its minimum value [Poletini, 2004].

Global weighting

Global weighting schemes [Manning et al., 2008] work on the term specificity, transforming the TDM terms $c_{i,j}$ independently for each term t_i according to its occurrence in each document d_j , which is called document frequency (df). Global weighting is based on the principle that the less occurrences a term has in the document collection, the more discriminant will it be.

- **Inverse Document Frequency:** Measures the relevance of a term t_i in a database. The more relevant a term is, the lower its inverse document frequency (idf) is, meaning that a term appearing in every document should have a value close to 0.
- **Probabilistic Inverse Document Frequency:** A modification of idf that spreads the possible values from $-\infty$ for terms t_i that appear in every document d_j , to 0 for terms t_i that appear in half the documents, to $\log(n - 1)$ for terms that appear in a single document.
- **Global Frequency Inverse Document Frequency:** Compares the global frequency of a term gf_i to its document frequency df_i , favoring terms that appear in smaller amount of documents for equal global frequency.
- **Entropy:** The odd one in the lot, entropy is an information theory concept that gives higher value to less frequent terms, as they are capable of providing more information than common terms. If a term t_i appears in a single document d_j its entropy will be the highest possible, and it will be 0 if it appears in all documents.

Global weighting techniques can also be used as a way of detecting stopwords ($idf \cong 0$), providing an alternative way to defining them at the pre-processing step. Even so, as applying stopword removal in the pre-processing step improves the weighting process efficiency, global weighting techniques are not commonly used for the stopwords detection task.

Querying in Genre Classification

Because we are applying a weighting scheme to the TDM, it follows that the same weighting must be applied to the query so as to be consistent. But, as the query is a single vector representation of the input text to be matched, there global weighting applied is the one obtained from the source database:

$$wq_i = lq_i * g_i$$

Where lq_i is the local weight of term t_i in the query q , and g_i is the global weight of term t_i in the documents database.

Finally, when particularizing the information retrieval task to the genre classification task we intend to use, the text to be synthesized by the expressive TTS system will become the query, with the documents database being all the expressive documents available that we can use for the purpose of genre classification.

2.4.3 Discrete genre modeling

Now that we have defined how we can obtain the different weighted term-document matrices and how to define the classification query, we have to start considering how to handle the information in order to obtain the classification results. For this task there are numerous approaches depending on the type of data to be processed [Baeza-Yates et al., 1999], but once again we will just consider unstructured text classification models. **Vector Space Modeling:** This model represents both documents and query as vectors in an m -dimensional space of terms, where m is the number of terms t_i in the vocabulary. These vectors are easily obtained from the WTDM matrix as the vector for document d_j will be the j -th column of the matrix. As such we start to consider the WTDM matrix the association of vectors

$$WTDM = [\vec{d}_1 \vec{d}_2 \vec{d}_3 \dots \vec{d}_n]$$

and

$$\vec{d}_j = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ w_{3,j} \\ \vdots \\ w_{m,j} \end{bmatrix} \quad \text{and} \quad \vec{w}q = \begin{bmatrix} wq_1 \\ wq_2 \\ wq_3 \\ \vdots \\ wq_m \end{bmatrix}$$

The results of the query $\vec{w}q$ will be obtained from a similarity measure between all documents $\vec{d}_j$ and the query itself. Regarding the similarity measures, there are many different ways of obtaining the correlation between vectors, but typically the cosine distance is used:

$$sim(\vec{d}_j, \vec{w}q) = \cos\theta = \frac{\vec{d}_j \cdot \vec{w}q}{|\vec{d}_j| \times |\vec{w}q|}$$

Vector space modeling, opposite to Boolean modeling, returns a value from 0 to 1, which is the correlation degree between our query and our database documents, which allows us to establish a ranking of similarities and with it train classifiers to be used for our intended final task, genre classification. Finally, the classification results can be used directly as a way of automatically selecting the correct expressiveness in the complete expressive TTS system proposed in this document.

2.4.4 Continuous vector space modeling

The discrete genre modeling described in the previous section is based on the assumption that all the terms of the vocabulary lists of the documents do not have any relationship to each other, which is ultimately false as language has semantic relationships that we should be able to model [Tonta and Darvish, 2010; Klein et al., 2011]. With that consideration in mind Latent Semantic Analysis (LSA) [Deerwester et al., 1990; Landauer et al., 2013], nowadays referred to as Continuous Vector Space Modeling [Krishnamurthy and Mitchell, 2013; Andreas and Ghahramani, 2013] was born. LSA aims to exploit the relationships between terms t_i and documents d_j by transforming them into an alternate "semantic" vector space, where both terms and documents are described by vectors of similar dimensionality and directions, enabling direct comparisons [Olmos et al., 2013; Cosma and Joy, 2012]. Typically this step is done by means of Singular Value Decomposition (SVD) [Golub and Reinsch, 1970; Henry et al., 2010], through which the latent semantic structure of the WTDM is shown.

Figure 2.9: Schematic of the SVD decomposition of the WTDM, from [Echeverry-Correa, 2015].

The objective of applying SVD to the weighted term-document matrix is to decompose it in three different matrices T , S , D^T as seen in figure 2.9 where T is the new **term matrix**, obtained as

$WTDM \cdot WTDM^T$, D is the new document matrix obtained as $WTDM^T \cdot WTDM$, and finally the result of the singular value decomposition, S , a rectangular matrix whose diagonal captures the singular values λ_i of the WTDM.

The singular values λ_i are known to capture the relationship between all terms and documents of the documents database, but we must remember that we are looking at a genre classification task where it is safe to assume that the number of terms will be much higher than the number of genres to be classified. Fortunately, it has been shown that LSA can be used also as a dimensionality reduction tool [Landauer and Dumais, 1997], where we can truncate the SVD matrix in order to obtain a new representation of the WTDM that still approximately represents the original matrix (figure 2.10), in a similar concept to what principal component analysis [Abdi and Williams, 2010] aims to do. The dimensionality reduction not only makes the system much more computationally efficient, but it also helps reduce the noise introduced naturally by the excess of terms in the vocabulary list.

Figure 2.10: Schematic of the truncated SVD, from [Echeverry-Correa, 2015].

Where:

$$WTDM = \hat{T} \cdot \hat{S} \cdot \hat{D}^T$$

Because we know that both the term and document matrices T and D are orthonormal, we can derive that:

$$\hat{D} = WTDM^T \cdot \hat{T} \cdot \hat{S}^{-1} \quad (2.2)$$

Finally, by reducing the dimensionality of the query by considering the input text as a document

(equation 2.2 [Berry et al., 1995]), we should be able to obtain a vector space representation in SVD of the desired text that can be comparable between documents because of the shared structure [Goyal et al., 2013; Blunsom et al., 2013]. This would allow us to pursue the rest of the tasks proposed for objective 1.3.3: we could obtain the SVD of a number of representative texts against which we can obtain a distance and with it a continuous genre classification, or we could even directly use the SVD directly as the lexical features of the synthesis training models.

SVD, VSM and LSA have been used successfully as the backbone of an unsupervised text to speech synthesis system capable of automatically learning grapheme to phoneme correspondences [Watts, 2012], and have seen a resurgence lately to the point that workshops specific to the technique are happening ².

2.5 Expressive TTS

2.5.1 Introduction

Current speech synthesis systems, whether we are talking about unit selection or HMM-based systems, can provide very good naturalness and intelligibility when synthesizing read speech regardless of the technology [Barra-Chicote et al., 2010; Barra-Chicote, 2011] which is ideal for neutral speech interfaces that do not need to engage in a direct conversation with the user. On the other hand, applications such as dialog systems [Lutfi et al., 2013] or virtual characters, where simulating a more human-like behavior is necessary, a neutral speech synthesis does not live up to the task. Imbuing the synthetic speech with expressive features (e.g. emotions, speaking styles...) is the role of expressive speech synthesis.

Due to the sheer amount of possible expressiveness, recording complete databases that cover all of them is unthinkable, making unit selection based systems fall behind in terms of scalability, although they are definitely capable of producing expressive speech [Erro et al., 2010b; Adell et al., 2010; Andersson et al., 2010; Adell et al., 2012]. On the other hand, HMM-based systems, because of their parametric nature, can be easily adapted through speaker adaptation techniques and can be successfully used for this task, and have been proven to provide significant improvements in perceived speech quality [Yamagishi et al., 2005].

²2nd Workshop on Continuous Vector Space Models and their Compositionality - <https://sites.google.com/site/cvscworkshop2014/>

One of the biggest problems of expressive speech synthesis is data acquisition. As human expressiveness is not a discrete space but a continuous one as the expressive strength and nuances vary greatly not only from person to person but from utterance to utterance for the same person. This problem can be focused on from different approaches: lexical analysis [Andersson et al., 2012] for correctly classifying the available data and training more precise systems or acoustic analysis. For acoustic analysis several aspects have been considered such as expressiveness detection [Lorenzo-Trueba et al., 2012a; Schuller et al., 2010; El Ayadi et al., 2011], expressiveness production [Obin et al., 2011; Raitio et al., 2013], expressive intensity control [Picart et al., 2011; Nose and Kobayashi, 2012] or expressiveness transplantation [Chen et al., 2012; Latorre et al., 2012].

The present PhD thesis is focused both in the expressive intensity control and the expressiveness transplantation field, as we want to propose the design of a system capable of controlling and transplanting the paralinguistic features of speech, as well as integrating it into a complete text-to-speech system as was explained in section 1.3.

2.5.2 Rule-based vs. Corpus-based Expressive TTS

As happened with the first approaches to speech synthesis with formant speech synthesis, the first approaches to expressive speech synthesis were rule-based approaches [Murray and Arnott, 1995]. This approach aims to modify the produced synthetic speech according to certain mathematical rules [Theune et al., 2006] or expert-knowledge rules [Murray et al., 2000] such as:

- Anger has a raised pitch contour and pitch range. With increased "laryngealisation". Speech rate is increased.
- Sadness has a lowered pitch contour, further lowered using Pitch again to give a bit of "laryngealisation" in the voice. Speech rate is slowed down, though not by too much.

Other rule-based approaches have focused on rules concerning unit-selection systems and the needed prosodic and phonological strategies to produce expressive speech [Pitrelli et al., 2006]. In general, these approaches tend to be successful in transmitting the expressiveness, but because they are based on rules and direct acoustic manipulation, they tend to introduce artifacts and to produce relevant decreases in speech quality [Zovato et al., 2004]. The biggest advantage of these approaches though is that they rely on expressive data only to define the transformation rules, but at synthesis time they morph non-expressive data, being more versatile than their corpus-driven

counterparts, which is why they are being studied even nowadays for languages and applications where data is sparse [Narupiyakul et al., 2013; Singh and Verma, 2013].

The alternative to rule-based expressive synthesis is the corpus-based expressive synthesis. This approach focuses on either learning from real expressive data if using HMM-based approaches [Yamagishi et al., 2004; Eide et al., 2004; Wu et al., 2006], or on plainly using expressive corpus in efficient ways for the unit selection process [Bulut et al., 2002]. These approaches tend to be able to provide higher quality speech while also providing the same expressive levels as rule-based approaches, with the disadvantage that they rely heavily on expressive data, not being able to imbue expressiveness in voices but to model and produce it by imitation. Even so, the advantages in speech quality and the modeling versatility provided by the corpus-driven approaches make them the main target of current expressive TTS systems [Charfuelan and Steiner, 2013; Eyben et al., 2012].

The system introduced in this PhD thesis is enclosed in the corpus-based approaches, but with the particularity that we intend to design a technique capable of not relying on having expressive data for all the speakers, but just for a reference expressiveness that will be transplantable to any desired target, effectively combining the advantages of rule-based approaches with corpus-based approaches.

2.5.3 Speaker Dependent vs. Speaker Independent Techniques

Scalability has been mentioned numerous times as one of the biggest limitations for expressive TTS systems, so considering whether an expressive speech synthesis technique is speaker dependent or independent is important depending on the intended application. A speaker dependent technique is only capable of producing speech for the speakers available as training data [Yamagishi et al., 2004; Bulut et al., 2002], combined with the intention of producing multiple expressiveness means that unless the training database is huge we cannot provide too many options. Unit selection systems, while capable of making use of rule-based approaches, are by default speaker dependent techniques, because mixing natural voice sample from different speakers is not naturally sounding.

On the other hand, speaker independent expressive synthesis techniques are meant to be able to produce the desired expressiveness in situations where no previous expressive data was available for the intended target speaker. This is of course doable with some of the rule-based approaches introduced in the previous section [Murray et al., 2000; Zovato et al., 2004], mainly with the

formant or HMM-based techniques. Another possibility for implementing speaker independent approaches is parting from a speaker dependent one and finally using a voice or speaker conversion technique of those introduced in section 2.2.4. The final possibility lies in exploiting the adaptability and versatility of HMM-based techniques in order to design transplantation algorithms that learn information from speaker dependent techniques and apply them in an speaker independent fashion [Chen et al., 2012], as introduced in section 2.5.4 and proposed as our second objective.

2.5.4 Expressiveness Transplantation Methodologies

Expressiveness transplantation methodologies can be defined as the procedures that allow the modification of a synthetic speech model to incorporate expressive information learned from other speaker models while maintaining the identity of the original speaker as much as possible. By this definition it follows that transplantation is a field of study that aims to solve one of the biggest problems in expressive speech synthesis: scalability.

A successful transplantation system should be capable of learning the paralinguistic nuances of the desired expressive speech model and then convert the target speaker model into a target speaker expressive model. But, as we would want the speaker identity to be maintained as much as possible, there are limitations on how much the features of the target speaker model should be modified [Togneri and Pullella, 2011]. In the end, systems have to reach a compromise between transplanted expressive strength and identifiability, synthetic speech quality and target speaker similarity.

A very successful transplantation method that has been introduced lately by Toshiba's Cambridge Research Laboratory is capable of producing highly expressive and high quality voices for automatic audiobook synthesis [Chen et al., 2012; Latorre et al., 2012]. This system is based on CAT, an adaptation based projective technology capable only of producing speaker models based on linear combinations of the training speaker models. This means that if we want to produce voices for a speaker with new expressive adaptation, the result will be the combination of the speaker training data that fits more closely the desired expressiveness, producing an expressive model for the target speaker. The main advantage of this approach is that as the produced model is always a combination of pre-existing training models, the process is extremely robust, outputting very high quality speech [Yanagisawa et al., 2013]. On the other hand, the level of expressive strength or speaker similarity cannot be guaranteed as the transplantation reach is very constrained.

Another approach to expressiveness transplantation is the use of rules to directly modify the synthesis models as introduced in section 2.5.2. This approach is theoretically capable of imbuing any expressiveness on any target speaker as long as we know the correct rules. In reality this approaches, while usually capable of providing expressive strength controllability and reasonably good recognition rates [Zovato et al., 2004; Takeda et al., 2013], speech quality and speaker similarity degradation tends to be a problem.

The second objective of this thesis (section 1.3.2) is to design an expressiveness transplantation method that considers the best of both previously mentioned approaches: using adaptation to lessen speech quality degradation while using the adaptation functions as pseudo-rules for modifying the speaker models. As a result we expect to design a method capable of controlling expressive strength while reasonably maintaining speech quality and speaker identifiability when compared to non-transplanted expressive synthetic speech [Lorenzo-Trueba et al., 2013b,a].

2.5.5 Cross-language Expressive TTS

Cross-lingual processing is always inherently difficult because of the inter-language differences that always exist. Traditional approaches rely on phonetic mapping between the source and target languages at different levels in order to do a correspondence between what is done in the source language in hopes that if the same is done for the destination language, the results will be acceptable.

The first attempts at phonetic mapping relied on monophones or triphones without contextual information in order to establish direct relationships between languages [Wu et al., 2008; Liang and Dines, 2011; Qian et al., 2011], and evolved to using contextual information by exploiting the decision tree structures by measuring distances between tree nodes of both languages [Wu et al., 2009]. Different approaches to the task tend to rely on having bilingual data for a single speaker and relying on state mapping and factor analysis techniques to extrapolate the knowledge that can be learned from the one speaker [Yoshimura et al., 2013; Zen et al., 2010].

Typical applications of cross-lingual technologies include adapting speaker models between different dialects, accents or variants of languages [Toman et al., 2013; Toman and Pucher, 2013], and we believe that we can consider different expressiveness or speaking styles the same way as dialects or accents, and use our proposed transplantation technique combined with cross-language technologies with the purpose of transplanting paralinguistic features between languages.

2.6 Expressive Corpora

When planing to work with speech processing technologies, using the correct corpora is imperative for obtaining the intended results. In particular for our task of designing an expressive text-to-speech systems, we must make sure that the expressive corpus we use fulfills not only the obvious speech and text quality requirements, but also that the different expressiveness are correctly defined and labeled. We also want our databases to be in Spanish initially, as most of our system will be developed as a Spanish TTS, but we can also learn from studying foreign language databases.

2.6.1 Emotional Corpora

We want our emotional corpora to cover at least the basic emotions defined in section 2.1.1 of this document: anger, happiness, sadness, surprise and also to consider the neutral emotion as a reference. There are large amounts of emotional speech databases in many languages [Burkhardt et al., 2005; Engberg et al., 1997], and there is a good study of what is required for developing a good emotional speech database [Douglas-Cowie et al., 2003]. But fortunately, previous studies in Spanish emotional speech introduced very high quality databases that fulfill all our needs.

Spanish Emotional Speech and Spanish Expressive Voices

Spanish Emotional Speech [Montero et al., 1998] and Spanish Expressive Voices [Barra-Chicote et al., 2008] (SES and SEV) are two acted emotional speech databases featuring a total of 2 male speakers and 1 female speaker. In SEV, which consists of the female speaker and one of the males, the 6 basic emotions are represented: Happiness, sadness, cold anger, fear, disgust, surprise and hot anger, with the addition of neutral voices too. SES only considers a smaller subset of them, namely happiness, anger, sadness, surprise and the neutral voice.

2.6.2 Speaking Styles Corpora

Speaking styles corpora suffer from the fact that defining what is a speaking style itself is a problem. Because of that, what we want from a speaking styles database is a clear way of distinguishing between the different speaking styles included. Probably the biggest problem in

defining a speaking style is that in common speech they tend to be extremely variable, and that is why most of the studies around speaking styles [Benjamin Picart and Drugman, 2013; Obin et al., 2011; Moreno-Sandoval et al., 2005] focus on professional speakers and their associated styles: news broadcast, sports commentaries, and so on or in different articulation styles such as breathy speaking, whispering or shouting [Picart et al., 2014; Raitio et al., 2013; Shozakai and Nagino, 2004].

For our proposed task we want to focus not on the articulation of speaking styles but more on speaking styles themselves both at linguistic and paralinguistic level. Consequently we are more interested on following an approach similar to the professional speaker and media styles route.

C-ORAL-Rom Database

C-ORAL-Rom [Moreno-Sandoval et al., 2005] is a multi-language and multi-style database covering a wide spectrum of formal and informal speaking styles. The languages included are French, Italian, Portuguese and Spanish, with styles ranging from formal to informal, extracted either from the media, telephonic conversations or natural speaking. For the task considered in this document we would only consider the Spanish formal media styles: news broadcasts, sports, meteorological reports, reportages, talk shows, scientific press and interviews. These data, having been extracted from media broadcasts of different stations presents a substantial variability in the recording environments and a high number of speakers (124). This implies that the database could not be the best source of information for learning paralinguistic features because the variances in the speaking styles between speakers of the same category are potentially big.

Glissando Database

Glissando [Garrido et al., 2013] is a Catalan-Spanish corpus that aims to provide a database for prosodic studies. The corpus is separated into two data-sets, news and dialogs. Dialogs are also separable into both conversational speech and task-oriented dialogs. The complete database features 28 speakers per language and about 25 hours of speech, with both transcriptions and prosodic annotations. In the end, Glissando database does not provide enough speaking styles variations for our needs, also suffering from being multi-speaker..

IRCAM Database

IRCAM [Obin et al., 2011] is a French multi-speaker multi-style database featuring 4 formal styles: news, live sports, political speeches and mass sermons. For each style it is guaranteed that multiple speakers and multiple environments are recorded, adding completeness to the database. This database proves that it is possible to define a completely separable subset of formal speaking styles even by using only the basic prosodic features of fundamental frequency and speaking rhythm, something that will be useful for the design process of the database explained in section 5.2.

2.7 Evaluation Methods in TTS

Evaluation is the most critical part of the design of any system, as thorough testing is the only way of proving it is capable of doing what it was designed for. It is also necessary to keep evaluating, at least informally, the evolutions of the system by means of objective measures. This can reassure us that progress in the correct direction is being made. In this section we want to analyze a few techniques we believe will be useful for the continuous evaluation of our planned tasks and objectives.

2.7.1 Latin Square

A Latin square of size n is a mathematical description of a n by n matrix in which each value occurs exactly once per row and per column (table 2.1), which makes it extremely useful as a tool for designing tests that cover a full spectrum of possibilities (i.e. systems) with as little test iterations as possible (n) [Gao, 2005]. This makes for extremely efficient testing environments at the cost of versatility because it aims to reduce all possible interactions between iterations of the test.

1	2	3	4	5
2	1	4	5	3
3	5	1	2	4
4	3	5	1	2
5	4	2	3	1

Table 2.1: Example of a 5 by 5 Latin square

This testing strategy has been used for a long time in medical treatment evaluations in order to use the minimum number of patients possible, and we believe it will prove extremely useful for designing the evaluation tests of our speech synthesis systems by minimizing the subjects that will have to undertake the perceptual tests in order for us to obtain satisfactory results [King and Karaiskos, 2013].

2.7.2 Perceptual Evaluations

In a perceptual evaluation the subject is presented with a sensory stimuli and then to answer a few questions about it. Traditionally speaking they have been used as a tool to measure human senses, and as such they are the only reliable way to test the results of a speech synthesis system. There are studies about objective quality measures [Quackenbush et al., 1988] that measure the signal to noise ratio [Ma and Loizou, 2011], artifact rates, reverberations [Kokkinakis and Loizou, 2011] and so on, but there is not a complete correlation between objective and subjective tests. So ultimately it is necessary to carry out a perceptual evaluation of the system if we want to know what a human thinks about it [Karhila et al., 2013].

Mean Opinion Score

Mean Opinion Score (MOS) [Viswanathan and Viswanathan, 2005] is a test that has been used for a long time as a perceptual quality measure of speech, originally designed as a way to measure the speech quality of telephone networks as defined by the ITU-T recommendation P.800³. In these tests the subject is asked to listen to a sound sample, and rate speech quality as they perceive it according to the scale in table 2.2. The final MOS value of a system then is the average of the rating of all the users that took the test.

MOS	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2.2: Mean Opinion Score value association.

³<http://www.itu.int/rec/T-REC-P.800-199608-I/en>

Preference Tests

Another approach to perceptual evaluation is presenting the listener with a pair of samples and asking whether they prefer one or the other. This approach is common when we want to compare systems that are similar between them but with variation in some conditions [Hecker and Williams, 2005; McHugh, 1976; King et al., 2008], and are a reliable way to obtain statistically relevant preference measures that are more separable than the traditional MOS evaluations. Additional advantages of preference tests is that it is generally easier to answer to a preference test than to rank a sample, making the time that must be invested in answering the test significantly shorter, which is always an advantage when carrying out perceptual evaluations.

2.7.3 Statistical Significance

Statistical significance displays the probability that a success was not caused by chance alone but that there could be a cause-effect relationship that justifies it [Steel et al., 1960]. Statistical significance and statistical hypothesis testing play an extremely important role in the evaluation of test results, as it is imperative to validate whether an obtained result is due to statistical variance or can be assumed to be statistically relevant. There are many different statistical significance tests that can be applied according to the task, but for our evaluation tasks in which we want to validate the significance between different evaluations of a single variable, paired difference tests are the norm.

Wilcoxon Signed-Rank Test

Wilcoxon signed-rank test [Hollander et al., 2013] is a paired difference test that can be used to verify the statistical significance when comparing two repeated measurements of a variable and to assess whether the mean values ranks differ outside statistical variation. The data must fulfill the following assumptions:

1. Data series to be tested must be paired and come from the same population.
2. Each pair should be chosen randomly and independently.
3. The data should be measured in an ordinal scale.

Fortunately, the results of a MOS test fulfill the assumptions, so Wilcoxon test will be a powerful tool for us to validate the statistical significance of our evaluation results.

Chi-squared test

Chi-squared test [Balakrishnan et al., 2013] defines any statistical test in which the distribution is a chi-squared distribution when the null hypothesis is assumed to be true. The distribution being chi-squared (with k degrees of freedom) means that it is the sum of squares of k independent normal random variables, a feature that is commonly used as a statistical independence test and in the definition of confidence intervals.

This validation tool will be very useful for defining the confidence intervals of all our functions, especially for the preference test validation as the Chi-squared test is a better fit than the Wilcoxon test.

Boxplots

Boxplots [Williamson et al., 1989] are graphical representations of numerical data through their quartiles, and also typically indicate variability ranges, maximum and minimum values, and much more. Boxplots allow us to verify easily but informally if the compared variables are statistically different and the variability ranks, as such it will be used extensively in our test validations together as a preface to the more formal Wilcoxon test or the Chi-squared test.

2.7.4 Proposed Evaluation Framework

For the present thesis we have considered a stable perceptual evaluation framework so as to be able to compare the different results. First of all, in order to eliminate learning bias, all evaluations when possible have been designed following the Latin Square strategy, which also contributed to maximize statistical significance while minimizing the need for evaluators. Regarding which evaluation scale was used, in general MOS was used for speech quality and perceived emotional strength evaluations. Also, when systems were evaluated independently we also utilized MOS, but when comparing systems we decided to use forced preference tests for the reasons described previously in this same section.

Speaker models were trained with three feature streams with their Δ and Δ^2 coefficients: logarithm of the fundamental frequency (1 coefficient), mel-cepstral analysis coefficients (MCEP, 60 coefficients) and aperiodicity bands (25 coefficients). For the synthesis part we used the STRAIGHT vocoder [Kawahara et al., 2001]. All evaluation samples were down-sampled from 48kHz to 16kHz.

In terms of the statistical significance of the evaluation results, a minimum confidence interval of 95% was fixed. Chi-squared test was used for the forced preference test analysis and Wilcoxon test for the MOS evaluations.

A complete description of the software and versions utilized for the present thesis can be seen in appendix B

Chapter 3

Controlling the Strength of Paralinguistic Properties of Speech

3.1 Introduction

The first step to obtaining a fully expressive TTS system is to be able to produce natural sounding expressive synthetic voices. Being able to control the different paralinguistic properties of the produced synthetic speech by controlling the acoustic features enables the system to control in a versatile fashion the produced expressive strength. This is very important for the objective of achieving a naturally sounding synthesizer because otherwise the output would be a monotonous speaking style unlike real human communication [Lutfi, 2013].

The groundwork for this method was carried out by Roberto in his PhD Thesis [Barra-Chicote, 2011]. His work focused on the study of emotional speech synthesis, particularly on interpolation and extrapolation of emotional speech. When we talk about the interpolation system in this chapter we refer to the implementation designed in the aforementioned thesis.

3.2 Adaptation based paralinguistic features control system

The most common method to achieve a paralinguistic strength control system is interpolation (see section 2.3.1). By obtaining a linear combination of different observations we can obtain a mixture that corresponds to our desired strength (e.g. combining a neutral and a happy obser-

vation, both at 50% we can obtain a half-happy observation). This simple approach is capable of providing reasonably good outputs, particularly when trying to produce outputs that do not extrapolate information from known data instead of interpolating (i.e. $\sum a_i \leq 1$). Likewise, this approach capabilities start decreasing rapidly when trying to emphasize the paralinguistic features as we will see in the results section of this chapter (section 3.3), which can become a problem if we want to produce fully expressive data despite the inherent smoothing caused by the HMM training process.

The approach proposed in this thesis is to apply adaptation techniques in order to better model the paralinguistic properties of the target speech and then work at model level instead of at observation level, which allows us to better compensate the Gaussian modeling smoothing by enhancing the expressive strength of the obtained models [Lorenzo-Trueba et al., 2012b]. This also allows us to generate speaker models with the desired expressiveness offline, which makes them more useable on real-time applications.

3.2.1 Control through adaptation

The developed system relies on learning the adaptation parameters (the rotation matrix $\mathbf{A}$ and the bias vector $\mathbf{b}$) that transform the source neutral model ($\mathbf{x}$) with the source expressive model ($\mathbf{y}$). The relationship between the variables is as follows:

$$\mathbf{y} = \mathbf{Ax} + \mathbf{b} = x + [(\mathbf{Ax} + \mathbf{b}) - x] \quad (3.1)$$

The adaptation function is learned using the CSMAPLR [Yamagishi et al., 2009] adaptation process introduced in section 2.3.2. This is the adaptation of choice throughout the whole system development because of its good capabilities when adapting expressive speaker models. Then, the paralinguistic features control itself is carried out by a parameter k . We can guarantee that the expressive controlled model $\mathbf{y}'$ is actually a linear interpolation between $\mathbf{x}$ and $\mathbf{y}$ by applying the following equation:

$$\mathbf{y}' = \mathbf{x} + [k(\mathbf{A} - I)\mathbf{x} + k\mathbf{b}] \quad (3.2)$$

Where I is the identity matrix. Finally, if we want the control to be applied as an adaptation itself we have to present it as a rotation matrix plus a bias vector pair, which we can do as follows:

$$\mathbf{y}' = [I + k(\mathbf{A} - I)] \mathbf{x} + k\mathbf{b} = [k\mathbf{A} + (1 - k)I] \mathbf{x} + k\mathbf{b} \quad (3.3)$$

The resulting equation 3.3 is the one that is implemented in the present system. Also, thanks to the rearranging of the coefficients, a control factor $k = 0$ results in the source neutral model $\mathbf{x}$, a control factor $0 < k < 1$ results in an interpolated model, a control factor $k = 1$ results in the source expressive model $\mathbf{y}$ and a control factor $k > 1$ results in an emphasized expressive model.

3.2.2 Proposed System

Figure 3.1: Proposed system for paralinguistic features control through adaptation.

The proposed system flowchart can be seen in figure 3.1 particularized for the emotion interpolation task, and it can be summarized in four steps:

1. Train an average voice model with Speaker Adaptive Training (SAT) [Anastasakos et al., 1997] with all the data available for the desired speaker in order to obtain a context-rich model for further steps.

2. Adapt the source neutral model $\mathbf{x}$ from the average voice model by means of CSMAPLR adaptation.
3. Learn the emotional transformation functions by adapting the different source emotions from the source neutral model to obtain the adaptation parameters $\mathbf{A}_{emo}$ and $\mathbf{b}_{emo}$.
4. Obtain the emotional controlled model $\mathbf{y}'$ for the desired control factor k making use of the equation defined in 3.3.

3.3 Perceptual Evaluation

The objective of this perceptual evaluation was to compare the paralinguistic features control capabilities of traditional interpolation-based control and the proposed adaptation-based control. The evaluation measured three variables: speech quality, emotional strength and emotion identification rates. Both systems were evaluated in the same perceptual test. Speaker similarity was not evaluated in the present evaluation but it was in the evaluation presented in section 3.3.5.

3.3.1 Evaluation Design

Four basic emotions (anger, happiness, sadness and surprise) learned from the female speaker in the SEV corpus (ROS) were evaluated in five different control factors ($k = 0, 0.5, 1.0, 1.5, 2$) for natural voice and both adaptation controlled and interpolation controlled systems. 30 minutes of speech were available for each emotion. A summary can be seen in table 3.1. This means a total of 39 systems (4 control ratios x 4 emotions x 2 evaluated systems + 2 neutral systems + 5 natural voice samples), which following the Latin-square approach [Gao, 2005] required 39 different utterances to be synthesized (or just reproduced in the case of natural voice) for all the systems. This meant that all systems were presented to the listeners in a random order without repetitions. Because the evaluation also considers emotion identification rates, the utterances were chosen to be devoid of any emotional context so that they do not influence the evaluation¹.

The evaluation was carried out by means of a web interface, where a single audio sample was presented to the listener without any additional information about the synthesized text or emotion. The samples could be played under demand by the listeners. First of all the listener was asked

¹Samples can be listened from <http://lorien.die.upm.es/~jaime.lorenzo/thesis/control.htm>

Systems	Emotions	Control Ratio (k)
Natural voice	Happiness	0.0
Adaptation	Sadness	0.5
Interpolation	Surprise	1.0
	Anger	1.5
		2.0

Table 3.1: Systems evaluated in the paralinguistic features control system.

to rank their perceived quality in a 5 points MOS scale (from very bad to very good), then they were asked to recognize the emotion conveyed by the utterance from an open list: happiness, anger, surprise, sadness, neutral or other. Finally perceived emotional strength was measured through a slider that ranged from “very low intensity” to “very high intensity”. Because emotional strength is measured through a continuous variable, the values had to be equalized along each individual listener (i.e. divide them by the maximum value selected by the evaluator). This equalization allows us to compare the results between evaluators while still allowing them to rate their perceived emotional strength in their own scale instead of relying on a fixed MOS scale.

A total of 39 subjects per evaluation was fixed in order to achieve the full coverage, meaning 39 samples per system, emotion and control ratio, for a total of 1521 utterances sampled.

3.3.2 Emotion Identification Rates

From tables 3.2 to 3.6 we can see the confusion matrices resulting from the perceptual test. Table 3.2 in particular is the confusion matrix for the natural voice, which will be the baseline model for comparison of recognition rates in the different techniques and control ratios.

Emotion identification rates for the interpolation-based system become equivalent to those of natural voice from ratios of $k=1.0$, which is reasonable considering that for $k=0.5$ the emotional strength is half of the natural voice. On the other hand for the adaptation-based system it is not until a $k=1.5$ that the recognition rates become comparable (see table 3.7).

It is also important to note how both negative emotions (sadness and anger) happen to be much more identifiable than the positive ones, who are significantly confused between them even for the natural voice. Also, by looking more closely to the natural speech, it was noted that surprise was acted as a ‘happy surprise’ which introduced significant confusion even in the natural voice recognition task (see table 3.2). As such it is safe to say that this is not due to problems in the

	Happiness	Anger	Surprise	Sadness	Neutral
Happiness	62.16	2.63	10.26	2.56	
Anger	2.70	68.42	5.13		5.41
Surprise	24.32		64.10		
Sadness		5.26	2.56	76.92	8.11
Neutral	8.11	10.53	2.56	17.95	78.38
Other	2.70	13.16	15.38	2.56	8.11

Table 3.2: Confusion matrix for the emotion identification rates for natural voice in the paralinguistic features control system evaluation in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.

Happiness					Surprise			
	0.50	1.00	1.50	2.00	0.50	1.00	1.50	2.00
Happiness	10.53	37.84	45.95	59.46	15.38	8.11	15.79	7.69
Anger			5.41	2.70	7.69	2.70		12.82
Surprise	5.26	18.92	21.62	21.62	2.56	43.24	31.58	46.15
Sadness	2.63		2.70		7.69	2.70	10.53	2.56
Neutral	78.95	40.54	21.62	13.51	56.41	35.14	28.95	15.38
Other	2.63	2.70	2.70	2.70	10.26	8.11	13.16	15.38

Table 3.3: Confusion matrix for the emotion identification rates of the adaptation-based system voice positive emotions in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.

Anger					Sadness				Neutral
	0.50	1.00	1.50	2.00	0.50	1.00	1.50	2.00	0.00
Happiness			2.63						2.63
Anger	13.51	48.72	73.68	92.31	7.69				7.89
Surprise	2.70	2.56							2.63
Sadness	2.70		5.26		10.26	42.11	82.05	94.74	
Neutral	72.97	41.03	13.16		76.92	55.26	15.38	5.26	78.95
Other	8.11	7.69	5.26	7.69	5.13	2.63	2.56		7.89

Table 3.4: Confusion matrix for the emotion identification rates of the adaptation-based system voice negative emotions in percentage. Rows show the chosen emotion and columns the actual emotion.

expressive strength techniques but because of peculiarities in the actor's happy emotions that makes them more difficult to distinguish.

Finally in table 3.7 we can see the average identification results across all emotions. There it is clear how the proposed adaptation-based system provides increasing recognition rates for higher

Happiness					Surprise			
	0.50	1.00	1.50	2.00	0.50	1.00	1.50	2.00
Happiness	18.42	56.41	59.46	41.03	2.63	21.62	24.32	13.16
Anger	2.63					2.70	2.70	2.63
Surprise	18.42	12.82	8.11	17.95	26.32	37.84	29.73	52.63
Sadness	2.63		2.70	2.56		8.11	10.81	5.26
Neutral	52.63	28.21	27.03	23.08	60.53	24.32	18.92	18.42
Other	5.26	2.56	2.70	15.38	10.53	5.41	13.51	7.89

Table 3.5: Confusion matrix for the emotion identification rates of the interpolation-based system voice positive emotions in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.

Anger					Sadness				Neutral
	0.50	1.00	1.50	2.00	0.50	1.00	1.50	2.00	0.00
Happiness			2.70	2.63		5.13			
Anger	16.22	62.16	62.16	55.26	5.26	2.56		2.63	2.56
Surprise	5.41		2.70		5.26		2.63	2.63	7.69
Sadness	16.22	2.70	2.70		28.95	64.10	65.79	81.58	10.26
Neutral	59.46	32.43	21.62	26.32	52.63	25.64	31.58	10.53	69.23
Other	2.70	2.70	8.11	15.79	7.89	2.56		2.63	10.26

Table 3.6: Confusion matrix for the emotion identification rates of the interpolation-based system voice negative emotions in absolute percentage values. Rows show the chosen emotion and columns the actual emotion.

control factors while the recognition rates for the interpolation-based system do not increase significantly since a $k=1.0$. We can also see how the proposed system provides better recognition rates than natural speech at the control factor of $k=2.0$ (73.2% for the adaptation-based system compared to the 67.9% of natural speech and the 57.6% of the interpolation-based system).

AVERAGE EIR	0.5	1.0	1.5	2.0
Natural	67.9	67.9	67.9	67.9
Adaptation	9.2	43.0	58.3	73.2
Interpolation	22.5	55.1	54.3	57.6

Table 3.7: Recognition rates in absolute percentage values averaged across all emotions.

3.3.3 Speech Quality

In Figures 3.2, 3.3 and 3.4 we can see plots of the perceived speech quality obtained for the different emotions comparing the 3 systems (natural voice, adaptation-based and interpolation-based) in MOS scale.

Figure 3.2: Results of the speech quality evaluation for natural voice in MOS scale.

Figure 3.3: Results of the speech quality evaluation for the adaptation-based control system synthetic speech in MOS scale.

In table 3.8 we can see the averaged speech quality results for the 3 considered systems. There we can see how there is at least around a 1 point decrease of speech quality in the MOS scale for the best conditions ($k < 1$). Also it is apparent how speech quality for the interpolation-based

Figure 3.4: Results of the speech quality evaluation for interpolation-based control system synthetic speech in MOS scale.

system falls sharply for the $k=2.0$ control factor (2.83 for $k=1.5$ falls down to 2.37). This also happens for the adaptation-based system, where there is a significant drop in speech quality from 2.94 for $k=1.5$ down to 2.76, but this drop is also significantly smaller than in the previous case.

AVERAGE SQ	0.5	1.0	1.5	2.0
Natural	3.94	3.94	3.94	3.94
Adaptation	2.95	2.96	2.94	2.76
Interpolation	3.04	3.06	2.83	2.37

Table 3.8: Speech Quality results in MOS scale averaged across all emotions.

3.3.4 Emotional Strength

Finally, for the normalized results of the estimated emotional strength (Figure 3.5) it is easy to see there are rising trends together with an increase in control factor K for the values introduced by the subjects of the test. There is only one combination of system and control ratio that breaks this tendency and it is once again the interpolation-based control system with a factor of $k=2.0$, where once again the extreme quality loss of the synthetic voice hinders any significant recognition pattern. It is also noticeable how for both happiness and sadness even for a control factor of $k=2.0$ the perceived emotional strength is smaller than that estimated for the natural voice, possibly due to bias caused in the listeners by differences in speech quality.

Figure 3.5: Plots of the perceived emotional strength of the four emotions in the paralinguistic features control system evaluation in a 1 to 100 scale.

In order to study the linearity of the results, we have obtained their Pearson correlation coefficient with the results of a linear system. The results can be seen in table 3.9, and it is clear that the adaptation-based system is much closer to the linear system, being effectively linear in all 4 emotions with an average correlation coefficient of 0.97, compared to the 0.82 of the interpolation-based system.

Correlation	Anger	Happiness	Sadness	Surprise
Adaptation	0.99	1.00	0.99	0.91
Interpolation	0.88	0.71	0.84	0.83

Table 3.9: Pearson correlation coefficient between the perceived emotional strength results and the theoretical perfectly linear system.

3.3.5 Iberspeech 2012 Evaluation

Iberspeech 2012 placed an interesting opportunity to test our system in an emotional speech synthesis challenge [Montero and Barra-Chicote, 2012]. This challenge provided a subsection of the male speaker in SEV (joa), consisting of the 4 emotions (anger, happiness, sadness and surprise) and also the neutral voice for a total of 500 utterances. The challenge then required to synthesize the test set for the evaluation process itself. As part of the competition we presented an implementation of the proposed adaptation-based paralinguistic features control system proposed in this chapter, but with a fixed $k = 1.5$ [Lorenzo-Trueba et al., 2012c].

SYSTEM	Speech Quality	Emotional Strength	Speaker Similarity	Emotion Identification Rate
Natural speech	0.87	0.71	0.43	0.78
Traditional synthesis	0.57	0.53	0.42	0.65
Anger	0.62	0.64	0.37	0.78
Happiness	0.57	0.61	0.48	0.65
Neutral	0.51	0.33	0.47	0.61
Sadness	0.49	0.52	0.41	0.65
Surprise	0.65	0.57	0.38	0.50
Adaptation	0.53	0.54	0.39	0.63
Anger	0.61	0.67	0.36	0.82
Happiness	0.53	0.59	0.40	0.34
Neutral	0.55	0.31	0.36	0.76
Sadness	0.49	0.56	0.32	0.79
Surprise	0.51	0.59	0.51	0.37

Table 3.10: Results relative to natural speech for the adaptation-based expressive control system and traditional speech synthesis[Iberspeech2012, 2012]. Bold values represent interesting values.

In table 3.10 we can see the results of the proposed control system relative to the results of natural speech in the challenge. The most notable result is that the emotion enhancing caused by setting the control factor to $k = 1.5$ significantly degraded the recognition of both happiness and surprise. Mainly because the synthesized surprise is a 'happy surprise' that becomes indistinguishable from happiness itself. In the other hand, emotional strength does not increase significantly, but it is expected that a better fine-tuning of the control factor would help increase the achievable results. Going further in the comparison, we can see how the control system is capable of improving the results in anger and sadness, mainly thanks to an improved in produced emotional strength. It is also fair to say that we are comparing the controlled system against the winner of the challenge, our proposed traditional emotional speech synthesis system.

3.3.6 Controlling Emotions in a Spoken Conversational Agent

The proposed paralinguistic features control system has also been implemented and evaluated in a dialog system [Lutfi et al., 2013]. The system consisted in a HiFi-control spoken conversational agent, which was designed to provide adaptive capabilities in the dialog, providing emotions according to the context. The objective was to mitigate users' frustrations and improving their overall satisfaction just by adding emotional capabilities.

For the evaluation, only happiness, sadness and neutral voices were considered, with emotion control factors ranging between 0.5 to 1.0 for happiness and between 0.0 (i.e. neutral) and 0.5 for sadness. The difference is because informal evaluations shows that neutral voice, in an emotional context, is generally perceived as sad, and also that a full-fledged sad voice does not fit a conversational agent task.

Metric	BASEHIFI	NEMOHIFI
Performance	2.13(X)	2.41(X)
Response	2.46(X)	2.29(X)
Voice	2.67	3.14
Attitude	1.90	3.11
Naturalness	1.42	2.54
GSS	2.00	2.69

Table 3.11: Mean results in MOS scale comparing a traditional, neutral TTS system (BASEHIFI) and an emotional system using the proposed method (NEMOHIFI) extracted from [Lutfi et al., 2013]. X denotes not significant differences.

The results of the spoken conversational agent evaluation can be seen in table 3.11. The results clearly show how the adaptive emotional system making use of our emotional control system clearly outperforms a traditional system based on neutral read-speech voices. Concretely, in terms of voice quality, attitude, naturalness and global satisfaction score (GSS) the emotional system shows significantly better results, which more than justify the advantages of using a controllable emotional system that adjusts the produced emotion according to the situation.

3.4 Paralinguistic Features Control System Conclusions

The developed adaptation-based paralinguistic features control system has successfully fulfilled the objectives we had in mind. First of all the evaluation showed how the proposed system is capable of maintaining good speech quality results even for control factors of 1.5 or 2.0, something that interpolation-based techniques failed at. This directly allowed us to compensate the known smoothing effect introduced by HMM modeling, allowing us to provide expressive synthetic speech closer in quality to natural speech, with better speech quality and identification rates that were attainable with traditional paralinguistic control systems.

Concretely, the proposed adaptation-based system provides an average 73.2% emotion recognition rate, maxing at 94.7% for sadness, compared to the average 57.6% provided by the interpolation-based system (both for $k=2.0$). In terms of produced emotional strength, the proposed system also provides better performance when compared to traditional systems as it is capable of providing linear increases for control ratios where interpolation-based systems break down. Numerically speaking, the proposed system shows an average Pearson correlation coefficient of 0.97 with the ideal linear system, compared to the average of 0.82 of the interpolation-based system.

The proposed system has also been proven to be comparable to traditional non-controlled emotional speech synthesis in terms of speech quality and speaker similarity (0.57 and 0.42 respectively for the traditional emotional TTS and 0.53 and 0.39 respectively for the proposed controlled system relative to the results of natural voice). It was also evaluated as part of an emotionally adaptive spoken conversational agent, where it significantly improved the performance of a neutral conversational system (1.69 compared to 1.00 in global satisfaction, 1.54 to 0.42 in naturalness or 2.11 to 0.90 in attitude).

Additionally, the developed system became an integral part of the following systems, being used directly in the semi-continuous version of the proposed automatic expressive TTS system. Also it became a critical stepping stone to the paralinguistic information transplantation system, as it was an evolution of this system to take into account not only expressiveness but also speaker identity.

Chapter 4

Transplanting Paralinguistic Information between Speakers

4.1 Introduction

The second step in the development of the fully expressive TTS system is to be able to transplant the paralinguistic information into a new target speaker, which is a significant step up from what we achieved in chapter 3. The proposed method is thus capable of imbuing expressiveness into the desired target speaker even when no expressive data is available as long as we have a source model for the target expressiveness.

4.2 Proposed system

The proposed system is an evolution of the paralinguistic features control system that, by means of adaptation techniques, is capable of learning not only the paralinguistic information of the source model but also the target speaker identity. Then, by means of cascade adaptations both paralinguistic information and target speaker identity are combined so that it can be applied to our source model. This system was successfully evaluated mainly for emotional speech as we will see in section 4.3 in both male and female speakers.

4.2.1 Emotion Transplantation based on Cascade Adaptation

The concept of cascade transforms has been used previously in automatic speech recognition to adapt the background models both to the target speaker and noise at the same time [Seltzer and Acero, 2012]. The transplantation method we present is based on the same concept, but in this case we propose chaining transformations that model both the emotional source speaker and the target speaker to produce emotional speech synthesis models, as introduced in the preliminary version of the system [Lorenzo-Trueba et al., 2013a].

Figure 4.1: Schematic of the emotion transplantation method.

In figure 4.1 we can see the block diagram representation of the proposed emotion transplantation method. If we define the CSMPALR adaptation functions in terms of their rotation matrix ζ and bias vector ϵ :

$$\bar{\mu}_{emo} = \zeta_{emo}\mu_N + \epsilon_{emo} \quad (4.1)$$

$$\bar{\Sigma}_{emo} = \zeta_{emo}\Sigma_N\zeta_{emo}^T \quad (4.2)$$

$$\bar{\mu}_{spk} = \zeta_{spk}\mu_N + \epsilon_{spk} \quad (4.3)$$

$$\bar{\Sigma}_{spk} = \zeta_{spk}\Sigma_N\zeta_{spk}^T \quad (4.4)$$

Where $\bar{\mu}_{emo/spk}$ and $\bar{\Sigma}_{emo/spk}$ are the mean vectors and covariance matrices of the emotional source models and target speaker model respectively. Then, the model resulting from applying in cascade the emotion transforms and then the speaker transforms becomes:

$$\bar{\mu}_{tar} = \zeta_{spk}\zeta_{emo}\mu_N + \zeta_{spk}\epsilon_{emo} + \epsilon_{spk} \quad (4.5)$$

$$\bar{\Sigma}_{tar} = \zeta_{spk}\zeta_{emo}\Sigma_N\zeta_{emo}^T\zeta_{spk}^T \quad (4.6)$$

The resulting emotional target model will be able to produce emotional synthetic speech for the target speaker even if emotional training data for the target speaker is not available.

4.2.2 Proposed Emotion Transplantation Method

The proposed transplantation method can be summed up in three steps:

1. Adapt a neutral average voice from the average voice model to work as a reference emotion (Fig. 4.2.1).
2. Adapt the neutral target model and the emotional source models from the neutral average voice (Fig. 4.2.2).
3. Apply in cascade the emotion and speaker transforms to the neutral average voice. The results are the emotional target models (Fig. 4.2.3).

The average voice model is obtained by applying SAT with as much training data as possible, which allows us to obtain a very context-rich background model to work with. A robust and complete AVM will be capable of producing better speech quality at synthesis time even with sparse emotional adaptation data. Also, sharing a background model for all the adaptation functions

Figure 4.2: Step by step block diagram of the emotion transplantation method. The ellipses represent the speaker models and the arrows the adaptation transforms.

makes the cascade adaptation easier, because the context decision trees will be shared, making the adaptation functions immediately compatible between adapted models.

Adapting the neutral average voice (NAV) from the AVM is necessary because in the second step we have two objectives: on one hand we want to be able to learn the differences between the emotional source model and the NAV, effectively learning the nuances of the desired emotional speech. On the other hand we want to learn the difference between the target speaker speaking in a typical, neutral style and the NAV, thus learning the nuances of the target speaker identity. If both adaptation functions are not obtained from a common reference emotion, neutral speech in the present case, they will not learn the desired characteristics and the transplantation process would not be successful. Ideally, we want to have both data for the target emotion and reference emotion for the same speaker so the emotion adaptation function defines purely the emotional source model, but if that is not available we can assume that using an average of different speakers will show the relevant information of the emotions while lessening the identities of the speakers.

Finally, we apply in cascade the emotion and speaker transforms as defined previously (Fig. 4.2.3), obtaining the desired emotional target models. The produced emotional strength can be easily controlled by means of a transplantation control ratio to linearly scale the adaptation function. Neither the target speaker nor the target emotion data had to be present in the AVM, and in the presented evaluation (section 4.3) we prove it to be successful with as little as 5 minutes of target speaker speech data or 30 minutes of emotional speech, which is why the proposed transplantation method is a good way of providing scalability in expressive speech synthesis. Nonetheless, in order to provide a basis for comparison, we also propose an emotion transplantation baseline.

Transplantation baseline: Transplanting into an average emotion

The proposed emotion transplantation baseline joins all our emotional data into an average emotion model [Qin et al., 2006] to be used as the emotional source. This average, when transplanted into the neutral target model, can be expected to imbue an undefined emotion that removes the typical monotony in read speech models. This alternative can also be expected to provide higher quality speech when compared to transplanting a single emotion as the adaptation process for the average emotion can make use of much more data, thus giving more stable adaptation functions. This approach could be very useful when the task does not require us to synthesize any particular emotion or if we do not have enough emotional data to obtain good emotion transplantation quality.

Naturally, this alternative also presents numerous shortcomings: if there is a significant bias towards positive or negative data in the average emotion model, transplanting the average emotion could be the same as transplanting an emotion, resulting on unnatural synthesized utterances for opposite emotions such as producing happy speech for a sad text. Another expected problem is that the naturalness should be lower than transplanting the correct emotion for the text to be synthesized.

4.3 Evaluation

4.3.1 Preliminary Evaluations

During the development of the proposed paralinguistic features transplantation system a number of preliminary evaluations were carried out considering both emotions [Lorenzo-Trueba et al., 2013a] and speaking styles [Lorenzo-Trueba et al., 2013b] transplantation. These evaluations allowed us to continuously improve and iterate on the proposed system implementation [Lorenzo-Trueba et al., 2015].

The first evaluation we carried out was on speaking styles transplantation and is explained in detail in section A.1. This evaluation considered alternatives to the proposed cascade adaptation approach based on CSMAPLR (i.e. copying the prosody or the spectrum of the expressive models instead of adapting). The adaptation-based proposed system was implemented by us on

HTS ¹, making use of the Parent transform functionality it provides. The results showed that even if the alternative approaches were comparable to our proposed system for less expressive models, our adaptation-based approach was more stable and introduced less artifacts into the transplanted models. These results were more significant when considering highly expressive models or source and target speaker models with more variations, which we considered satisfactory and led us to continue with the present research.

After the first preliminary evaluation we streamlined the transplantation process, developing a set of scripts designed to automate it, and carried out a second one, our first emotion transplantation evaluation [Lorenzo-Trueba et al., 2013a]. This second evaluation measured the similarity, emotion recognition rates, speech quality and emotional strength of the proposed systems with different emotion transplantation control ratios, using neutral synthetic speech as baseline and natural voice as a top-line. Only the male source and target speakers of the introduced corpora (section 2.6) were used for this preliminary evaluation. The results for the emotion identification rates can be seen in table 4.1. The test showed how that the attainable similarity is comparable to natural speech, and that values of the transplantation control ratio that keep a good balance between speech quality and emotional strength range between 0.75 and 1.00. This was because higher ratios degraded the speech quality too much and lower ratios limited too much the emotion identification rates and the perceived emotional strength.

Metric	System	Anger	Happiness	Sadness	Surprise
Speech Quality	Proposed	67.3	62.8	60.4	62.8
Emotion Id.	Proposed	36.0	52.1	88.5	73.1
Emotional Strength	Proposed	65.2	69.1	67.7	66.4
Similarity	Proposed	84.6	94.5	86.2	67.8

Table 4.1: Summary of the results for the preliminary evaluation [Lorenzo-Trueba et al., 2013a] on emotion transplantation for speech quality, emotion identification, emotional strength and similarity to the target speaker for a transplantation control ratio of 1.00. Results are relative to those of the natural voice and showed in percentage values.

The results of both preliminary evaluations show the strength and limitations of the proposed transplantation system from a task-independent point of view. We saw how the system is capable of correctly transplanting the source expressiveness while maintaining natural speech levels of target speaker similarity even in abstract expressiveness identification evaluations. Thus, the proposed evaluation focused on analyzing the transplantation system in a more realistic emotional text synthesis task, where the synthetic voices ideally match the text to be synthesized.

¹HMM-based Speech Synthesis System (HTS) - <http://hts.sp.nitech.ac.jp/>

4.3.2 Evaluations Design

The goal of the perceptual evaluation presented here was to verify if the expressiveness was transplanted successfully in terms of naturalness, speech quality and emotional strength. Naturalness measure was done by means of forced preference tests, as they are very useful when we want to directly compare systems that are similar to each other but with variation in some conditions [King et al., 2008], which is why we chose this approach over a MOS scale evaluation. Three different evaluations were carried out: a first evaluation that compared the average emotional model transplantation against the traditional neutral synthetic voices in order to establish an emotion transplantation baseline. A second evaluation to compare the proposed emotion transplantation system against neutral voices so as to validate the usefulness of the presented methodology. A third evaluation comparing the average emotion transplantation with the proposed emotion transplantation system, aiming to rate the performance of the proposed system against the baseline.

Four emotions (anger, happiness, sadness and surprise) learned from the SEV corpus were transplanted into 3 male speakers and 3 female speakers, so for each test the total number of systems was 24. Following the Latin-square [Gao, 2005] approach this meant that we needed 24 different utterances to be synthesized for all the systems to be presented to the listeners in a random order without repetitions. In the end we decided for considering 24 utterances per evaluation and not multiples of it so as to keep the evaluation itself from becoming too long, lasting around 30 minutes in this implementation. In the test, two audio samples were presented to the listeners by means of a web interface: the transcription of the synthesized texts and the intended emotion to be transmitted. The samples could be played as many times as desired by the listeners. The synthesized texts, not present in the training data, were carefully designed by ourselves to present clear emotional context as it we wanted to evaluate the adequacy of the voice to the text. A sample utterance for each emotion can be seen in table 4.2 (and table 4.3 for the Spanish version). The listeners were asked their preference for which of the samples was more adequate to transmit the desired emotion, being forced to pick one or the other. Then they were asked to rate the utterances in the traditional 5 point MOS evaluation for both speech quality (from very bad to very good) and emotional strength (very low to very high). The evaluation for speech quality and emotional strength had to be answered for both samples regardless of the selected preference. In the case of the multiple choice secondary evaluation (subsection 4.3.4), the listeners were presented the transcribed synthesized text, the target emotion and a set of 5 utterances and were asked to select one sample out of the five that they believed better conveyed the target

emotion. The 5 utterances corresponded to the 5 emotions (anger, happiness, neutral, sadness, and surprise) and were randomized in order for each sample.

For the present evaluation we did not measure similarity or the emotional strength control capabilities as they were already studied in the preliminary evaluations, and the applied emotion transplantation control ratio was 1.00. Regarding the statistical significance of the results, for the preference test we applied the chi-squared criterion and for the speech quality and emotional strength MOS tests we applied the Wilcoxon Signed-Rank Test for a 95% confidence ratio. A minimum of 24 subjects per evaluation was fixed in order to guarantee the full coverage of the Latin square matrix, totaling at least 24 samples per speaker, emotion and evaluation².

Emotion	Utterance
Anger	I have just picked up my clothes from the laundry, they have been ruined!
Happiness	I am so happy! I passed the exam in the first try even though the subject was so hard.
Sadness	It has been two years since he passed away, we miss him dearly.
Surprise	You cannot believe who I just saw walking down the street!

Table 4.2: Sample emotional utterances used for the proposed transplantation system preference evaluations translated from Spanish .

Emotion	Utterance
Anger	Acabo de recoger mi ropa de la lavandería, ¡me la han estropeado!
Happiness	¡Qué contento estoy! Aprobé el examen a la primera aunque la asignatura era muy difícil.
Sadness	Hace dos años que falleció, la echamos mucho de menos.
Surprise	¡No te vas a creer a quién acabo de ver caminando por la calle!

Table 4.3: Sample emotional utterances used for the proposed transplantation system preference evaluations in Spanish .

Regarding the speaker models, the three considered systems (neutral, average and proposed) both share a common AVM trained with three feature streams with their Δ and Δ^2 coefficients: logarithm of the fundamental frequency (1 coefficient), mel-cepstral analysis coefficients (MCEP, 60 coefficients) and aperiodicity bands (25 coefficients). For the synthesis part we used the

²Samples can be listened from <http://lorien.die.upm.es/~jaime.lorenzo/thesis/transplantation.htm>

System	Neutral Source	Emotional Source	Neutral Target
Neutral	JOA/ROS	None	All 6 target speakers
Average	JOA/ROS	All 4 emotions	All 6 target speakers
Proposed	JOA/ROS	Target emotion	All 6 target speakers

Table 4.4: Source and target data used for training the adaptation functions for the 3 systems in the paralinguistic features control system evaluation.

Speaker	Gender	Emotions	Amount of data
JOA	male	A,H,S,U,N	2h 30min
ROS	female	A,H,S,U,N	2h 30min
UVD	male	N	2h
JEC	male	N	7 min
JLC	male	N	3 min
UEM	female	N	1h 45min
NAS	female	N	7 min
EMA	female	N	6 min

Table 4.5: Complete characterization of the speakers used in the emotion transplantation evaluations. In the emotions section, A stands for anger, H for happiness, S for sadness, U for surprise and N for neutral.

STRAIGHT vocoder [Kawahara et al., 2001] for all the systems. Training data can be seen in table 4.4 and 4.5. All emotions besides neutral consist of approximately 30 minutes of speech.

4.3.3 Evaluation 1: average emotion transplantation system vs. neutral system

The first test aims to establish the baseline by comparing the average emotion transplantation system and the neutral synthesis system. The first aspect to note is that the preference test results greatly favor the average system as we can see in table 4.6, with an average preference rate of 75%. The results for emotional strength also show that the average system is perceived generally as more expressive, with an average of 0.7 points more perceived strength than in the neutral system, especially in the angry and surprised cases. Finally, speech quality suffers a slight degradation of 0.2 points in average, being this category the only one with non significant differences (figure 4.3). All the boxplots shown in this paper follow the same structure: the boxes capture the ranges between the first and third quartile, with the median represented by the thick line and the whiskers showing the highest and lowest datum within 1.5 the inter-quartile range.

Figure 4.3: Boxplots for the results of both speech quality and emotional strength for the first transplantation evaluation. The blue bars represent the average emotion transplantation system and the red bars represent the neutral system.

Metric	System	Anger	Happiness	Sadness	Surprise	Average
Preference	Average	66%	87%	61%	84%	75%
	Neutral	34%	13%	39%	16%	25%
Speech Quality	Average	3.11	3.19	3.16	3.21(X)	3.17
	Neutral	3.38	3.38	3.38	3.38(X)	3.38
Emotional Strength	Average	3.37	3.33	2.93	3.34	3.24
	Neutral	2.57	2.57	2.57	2.57	2.57

Table 4.6: Results for the first transplantation evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant. Quality and strength are represented in MOS scale, preference in percentage.

4.3.4 Evaluation 2: proposed emotion transplantation system vs. neutral system

The second test looks to show the advantages of the proposed emotional target model when compared against the neutral target model. The results (table 4.7) show a clear preference for the transplanted system (average of 88%). This rate is higher than in the first evaluation, and reaches rates as high as 96% or 95% for happiness and surprise respectively. Emotional strength results prove an average increase of 1.2 points in the perceived emotional strength, especially again in the happy and surprised systems, clearly shown in the boxplots (figure 4.4). Speech quality suffers a slightly stronger degradation (an average of 0.4 points).

Figure 4.4: Boxplots for the results of both speech quality and emotional strength for the second transplantation evaluation. The green bars represent the proposed transplantation system and the red bars represent the neutral system.

Metric	System	Anger	Happiness	Sadness	Surprise	Average
Preference	Proposed	77%	96%	82%	95%	88%
	Neutral	23%	4%	18%	5%	12%
Speech Quality	Proposed	3.05(X)	2.81	2.70	3.05	2.90
	Neutral	3.30(X)	3.30	3.30	3.30	3.30
Emotional Strength	Proposed	3.03	3.61	3.26	3.73	3.41
	Neutral	2.17	2.17	2.17	2.17	2.17

Table 4.7: Results for the second transplantation evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant. Quality and strength are represented in MOS scale, preference in percentage.

Follow up to evaluation 2: proposed emotion transplantation system vs. neutral system multiple choice preference

In order to measure the capabilities of the proposed emotion transplantation system to convey the desired target emotion we carried out a follow up to the second test, where we asked the evaluators to pick the utterance they felt that conveyed more clearly the target emotion out of a pool of 5 (anger, happiness, neutral, sadness and surprise). The confusion matrix resulting the evaluation can be seen in table 4.8.

The results show how the evaluators significantly prefer the utterances that were transplanted with the desired emotion when compared to the rest of the utterances with an average preference

Preference	Anger	Happiness	Sadness	Surprise
Anger	67.9	2.6	14.2	3.4
Happiness	14.7	71.1	3.2	20.1
Sadness	1.3	2.0	70.3	3.4
Surprise	14.7	23.7	4.5	71.8
Neutral	1.3	0.7	7.7	1.3

Table 4.8: Confusion matrix in percentages for the multiple choice preference test considering all four transplanted emotions and neutral voice. Rows show the chosen preferred system and columns the actual transplanted emotion.

of around 70% for the intended one, which proves that the method is capable of adequately transplanting emotions. In particular we can see how there was almost no preference for the neutral utterances, even in the sad system where a neutral way of speaking could sometimes be acceptable. Also, the confusion between happiness and surprise makes sense, with both of them being positive emotions with high arousal that tend to produce higher pitched voices. The confusion of anger with happiness and surprise can be explained by the fact that sometimes anger is conveyed by a high arousal variation called hot anger. Sadness on the other hand is slightly confused with both cold anger and neutral, both low arousal states with slower speaking rhythm and lower pitch.

4.3.5 Evaluation 3: proposed emotion transplantation system vs. average emotion transplantation system

This evaluation aims to measure the advantages of the proposed transplantation system against the baseline average emotion transplantation system, which means measuring the advantages of transplanting the emotion associated to the text to be synthesized versus using an average standard emotion for all texts. In this case, the results clearly show a preference for the proposed transplantation system, with an average preference rate of 75% (table 4.9). Also, the perceived emotional strength results show an increase of 0.5 points when selecting the adequate emotion, stronger for the happy and surprised systems (boxplot 4.5), but also present in the angry and sad systems. For speech quality we see extremely close results, less than 0.1 points.

Figure 4.5: Boxplots for the results of both speech quality and emotional strength for the final transplantation evaluation. The green bars represent the proposed transplantation system and the blue bars represent the average emotion transplantation system.

Metric	System	Anger	Happiness	Sadness	Surprise	Average
Preference	Proposed	57%	80%	76%	84%	76%
	Average	43%	20%	24%	16%	24%
Speech Quality	Proposed	3.22(X)	3.06	3.07(X)	3.12	3.12(X)
	Average	3.13(X)	2.94	3.06(X)	2.97	3.03(X)
Emotional Strength	Proposed	3.10	3.24	3.12	3.35	3.20
	Average	2.88	2.67	2.50	2.66	2.68

Table 4.9: Results for the final transplantation evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant. Quality and strength are represented in MOS scale, preference in percentage.

4.4 Evaluation of the transplantation system in a real robotic application

In order to evaluate the improvements attainable by the proposed emotion transplantation method, we also designed an evaluation in a more realistic application: synthesizing the voice of a robotic laboratory assistant (figure 4.6). This robotic laboratory assistant, in its finished form, would be capable of moving around a laboratory interacting with the students and answering their questions or doing quizzes to evaluate their progress. For this particular evaluation, as we wanted to focus on how we could improve the robot’s interaction capabilities by adapting its way of speaking to the context, we only used the robot’s head in a ‘Wizard of Oz’ dialog scenario. Our role as the ‘Wizard of Oz’ was to replace the speech recognition system, starting the robot’s turns in the scripted dialog.

4.4.1 Robot description

We have developed a face with 5 degrees of freedom, 2 to control the mouth, one for each eyebrow and another one for closing both eyes, similar to [Rodriguez-Losada et al., 2008]. The eyes were completely static, the mouth could not be opened and both eyelids had to be moved together. The actuators are based on Futaba servomotors and a simple servo-controller board allows sending sequences of control bytes through a serial port from a portable PC on-board the robot. As shown in Figure 4.7, it is perfectly capable of showing basic emotions such as happiness, sadness, anger or surprise. We want to thank the ARABOT project members for their invaluable help on designing and building the robot.

Figure 4.6: Picture of the robotic assistant, Urbano.

Figure 4.7: Expressive faces for the robot in the following order: neutral, happiness, sadness, anger and surprise.

4.4.2 Robot evaluation design

So as to emulate a realistic scenario, we developed a total of 36 interaction scenarios (12 for happiness, 12 for sadness, 6 for anger and 6 for surprise) of which 6 would be carried on by each evaluator (2 for happiness, 2 for sadness, 1 for anger and 1 for surprise). A sample dialog translated to English can be seen in table 4.10 and in table 4.11 in Spanish. This dialog distribution was chosen because we believe that it is more common in a student-robotic assistant dialog to show happiness and sadness than anger or surprise. Regardless, all of them had to be present to fully test the proposed system. This evaluation was also designed with the Latin square principle in mind, so that the order of the emotions and dialogs was fully randomized. In each scenario a short dialog (between 1 and 3 interaction turns) would be played out between the robot and the evaluator, and it would be played twice (once for the neutral system and once for the proposed

Turn	Robotic Assistant	Student
1	Good morning, can I help you?	Yes, I wanted to measure this, but something is wrong.
2	Have you tried measuring with the second channel?	Oh! That was the problem! Thank you very much.
3	<i>Great! I was glad to be of help.</i>	

Table 4.10: Sample happiness dialog used in the robot evaluation, translated from Spanish. The italics text shows the emotional turns in the dialog.

Turn	Robotic Assistant	Student
1	Buenos días, ¿Puedo ayudarte en algo?	Sí, quería medir ésto, pero algo falla.
2	¿Has probado a medir en el segundo canal?	¡Oh, ése era el problema! Muchas gracias.
3	<i>¡Genial! Me alegro de haber sido de ayuda.</i>	

Table 4.11: Sample happiness dialog used in the robot evaluation in Spanish. The italics text shows the emotional turns in the dialog.

emotional transplantation system, in a random order). The robotic face would then adjust its expression to match that of the context regardless of the system that was being used, so that the only variation between the iteration of the dialogs was the synthetic speech. Also, to prevent the evaluation from becoming too long, we only considered one of our target speakers, UVD.

The evaluators, which were real students from UPM university, were explained that they were going to evaluate a robotic laboratory assistant that wanted to be capable of adapting its way of speaking according to the context. Then they were given the scripts for the dialogs and were told that each dialog would be played twice, once for each emotional system. After carrying out each scenario twice, the evaluator was asked to choose between both versions which one they felt was better according to the dialog that was played out in a forced preference test. If they were undecided, the scenarios could be carried out multiple times, but only the complete scenario.

4.4.3 Robot evaluation results

The results for the 24 students that took part on a first evaluation can be seen in table 4.12. There we see how the preference is heavily in favor of the proposed emotion transplantation system, which clearly shows that being able to provide a speech synthesis system capable of adapting its speaking style (emotions in this particular application) to the context can provide

Metric	System	Anger	Happiness	Sadness	Surprise
Preference	Proposed	77	83	54 (X)	88
	Neutral	23	17	46 (X)	13

Table 4.12: Results in percentage for the first preference evaluation with the robotic assistant. The X represents results that are not statistically significant.

Metric	System	Anger	Happiness	Sadness	Surprise
Preference	Proposed	83	83	46 (X)	83
	Neutral	17	17	54 (X)	17

Table 4.13: Results in percentage for the second preference evaluation with the robotic assistant. The X represents results that are not statistically significant.

significant increases in the satisfaction of the users of the dialog system. It is only sadness that did not provide significant improvements, and informal dialog with the evaluators that disliked it showed that the problem was that they felt that the sad system was out of place in the laboratory assistant. Careful inspection of the source sadness data showed that it presents a tearful tone, which adds an aspect to the sad emotion that makes it inadequate for this particular domain.

In an attempt to fix the problems with sadness we prepared a second evaluation in which sadness was transplanted with only a 0,5 transplantation control ratio so that the synthesized emotional strength was lessened, while the rest of the emotions remained at 1,0. In this second iteration of the system 12 students participated, giving results very similar to the previous iteration as we can see in table 4.13. Informal conversation once again showed that while less recognizable, the effects of tearfulness still made the transplanted sadness to feel out of place, showing that reducing the emotional strength is not enough to attenuated the tearfulness aspect from the sad models as they are conveyed by complex spectral cues that cannot be eliminated without removing also sadness itself.

The conclusion that can be drawn from the sadness results is coherent to what we saw in the evaluation of the spoken conversational agent that made use of the proposed paralinguistic features control system (see section 3.3.6). In a conversational agent, neutral voice is commonly considered as sad when compared to other expressive voices, and a full-fledged sad voice can be overwhelming for the user and not appropriate for the task. As such, it is reasonable to conclude that for any application involving robotic assistants or dialog systems it is better to use a neutral voice instead of sad voices.

4.5 Paralinguistic Information Transplantation System Conclusions

We have developed an emotion transplantation method capable of learning the paralinguistic nuances of any particular emotion in order to transplant them into a new target speaker for whom only neutral read speech recordings are available. This is done by means of chaining a pair of CSMAPLR adaptation functions, one that characterizes the target speaker identity and another that captures the paralinguistic characteristics of the desired emotion. Finally a triplet of perceptual evaluations were carried out.

For the perceptual evaluations, four emotions (anger, happiness, sadness and surprise) from a Spanish emotional database and six target speakers (three male and three female) were considered. A first evaluation compared in terms of naturalness, speech quality and emotional strength a baseline average emotion transplantation with traditional neutral read speech synthesis. The second evaluation compared the proposed emotion transplantation method with neutral speech synthesis, and the final evaluation compared the baseline with the proposed method. The results clearly proved how emotion transplantation greatly increases the perceived emotional strength of the synthesized utterances (up to 1.2 points for the proposed method), also showing a very significant increase in preference (an average of 88% for the proposed system) at a cost of only 0.4 points in speech quality. This results are enforced by the average 70% preference rates in the multiple-choice evaluation. The comparisons also showed how transplanting the correct emotion is a much better alternative than transplanting an average emotion, with significant increases of 0.5 points in perceived emotional strength and a 76% preference rate at a cost of less than 0.1 points in speech quality. We have also seen that there is no correlation between speaker gender or identity and the obtained results, while there is a strong correlation between the source neutral system speech quality and the transplanted quality, meaning that the proposed transplantation system is robust against speaker variability, and is capable of providing high quality emotional models if the source neutral speaker model is of high quality. The transplanted voices provided us with multiple emotional voices for the robots developed in the ARABOT and INAPRA projects, but by imbuing emotions on previously available speakers recorded in just a neutral speaking style.

After designing, developing and evaluating the paralinguistic features transplantation method, we applied it to different problems in order to test it under more situations such as speaking

styles transplantation or cross-lingual emotion transplantation. We also tried to develop a prediction system to detect potential artifacts in the transplantation process to reduce the quality degradation, and finally we attempted a different approach to transplantation in the shape of a MR-HMM based transplantation system. The studies are presented in detail in Annex A.

This additional work helped us set the foundations for a cross-lingual emotion transplantation system (section A.3), which enabled us to successfully convert the adaptation functions that convey emotional information from a source language (Spanish) to the desired target language (English) by means of a state mapping technique based on minimizing the KLD between neutral language models. This has enabled us to correctly modify the prosody and spectral components of the target speaker in the target language in the same fashion that we would have done in the source language, conveying then the desired emotion.

In an attempt to characterize completely the artifacts introduced by our proposed paralinguistic features transplantation system we developed a modification of the acoustic mapping method introduced in the cross-lingual system. The aim of this system was to attempt to predict whether any particular context will be corrupted by the transplantation function and thus produce an artifact (see annex A.4). This system by itself did not prove to be powerful enough to predict artifacts in the speaker models, but it showed a significant trend in which the bigger the distance between the transplanted contexts, the higher the probability of it being corrupted.

Also we developed an emotion transplantation system based in multi-factor multiple-regression HSMM synthesis (section A.5), which just by providing a control vector consisting in both speaker and emotion factors we can obtain a transplanted speech model. Segmental and mixed-nature emotions have been partially transplanted (48% for anger and 73% for happiness relative EIR to the source speaker). Prosodic emotions have been successfully transplanted. In the case of speaker similarity, we obtained an average preference value of 77% for the target speaker, proving that we are capable of maintaining the speaker similarity of the transplanted speaker. This initial work on MR-HMM helped us develop the continuous expressive TTS system that correctly synthesized speaking styles relying only in lexical information.

Chapter 5

Expressive TTS System: From Text Genre to Speaking Styles

5.1 Introduction

This final piece of the desired expressive TTS system is in charge of automatically predicting the most adequate speech synthesis model to synthesize a target utterance. The prediction was carried in a number different ways: a discrete prediction that only predicts the adequate speaking style and synthesizes accordingly, a semi-continuous prediction that together with the genre prediction obtains an intensity value that is used to control the synthesized expressive strength. We also consider a continuous approach that does not rely on genre prediction but on the CVSM vector as part of a Multiple-Regression HMM adaptation to implicitly model genre from the text replacing the prediction step.

5.2 Spanish Speaking Styles

The present system aims to combine genre prediction techniques and expressive speech synthesis. As such, we want a corpus that not only covers a number of speaking styles with a reasonable amount of text data, but also speech from a single speaker, so that the synthesis models provide high quality synthetic speech. With that objectives in mind we designed and recorded Spanish Speaking Styles (SSS). For the present database we focused not on the articulation of speaking

styles but more on speaking styles themselves both at linguistic and paralinguistic level. Consequently we are more interested on following an approach similar to the professional speaker and media styles route.

5.2.1 Database Design

The first aspect to define in the database is the selection of which speaking styles were to be recorded, as they had to be sufficiently representative so as to be identifiable but at the same time they should cover a large spectrum of the expressive map. Also, we had to decide whether we want those styles to be scripted or unscripted. Both the scripted and unscripted styles topics were defined for approximately 1 hour per style in mind. In the end we decided for two scripted speaking styles: broadcast news and political speech, and two unscripted styles: live sports broadcasting and interviews. They were chosen in order to cover a wide range of expressiveness, which translated into the F0-rhythm map that we can see in figure 5.1.

Figure 5.1: F0 vs. rhythm map of the 4 recorded speaking styles in SSS.

Broadcast news scripts were adapted from 'Agencia EFE' news agency. Political speech scripts were reproductions of three different speeches given by former Spanish presidents. Both news and political speeches were fully prompted and recorded in paragraphs. Live sport was recorded

Speaking Style	Train set	Test set
NEWS	102 utts, 1h3min	21 utts, 11min
INTERVIEW	332 utts, 45min	21 utts, 7min
SPORT	200 utts, 56min	21 utts, 7min
POLITICAL SPEECH	116 utts, 56min	21 utts, 11min
TOTAL	750 utts, 3h40min	84 utts, 36min

Table 5.1: Detailed description of the SSS database.

Speaking Style	Train set
NEWS	Más de 50 millones de muertos en todo el mundo ha causado en la última década el consumo de productos del tabaco, una adicción que enferma a las sociedades y amenaza a las economías.
INTERVIEW	Ya. Bueno pues vamos ya entonces al cuarto país en tu ranking de países favoritos, Estados Unidos. ¿Qué zona de Estados Unidos, de ese país que casi es un continente has visitado?
SPORT	Cesc, dentro del área se mete hasta la cocina y, gol, gol, gol gol gol gol gol. Nadie esperaba esto, gol de Silva. Qué gol desde atrás del canario.
SPEECH	Nunca he comprendido y sigo sin comprender, esa idea del centro, como el final imposible de un viaje interminable. Nadie nos tiene que enseñar el camino del centro.

Table 5.2: Sample test sentences from SSS. News and political speech samples are shortened for representation purposes.

as the speaker was broadcasting a football match played on a display while listening to real background noise recorded during another match. Interviews were fully unscripted and recorded continuously. The unscripted styles (interviews and sports) were hand labeled and manually segmented. For all four styles, three types of pauses were hand labeled: intra-sentence pauses, inter-sentence pauses and filled pauses for better prosody modeling. 21 utterances for each style were separated for the test set and were also recorded in a neutral speaking style for evaluation purposes. A summary of the database can be seen in table 5.1, and some test utterance samples in table 5.2

5.2.2 Recording Environment

The speaker was located inside an acoustically treated room equipped with a close-talk head-mounted microphone and a set of headphones for listening to instructions and also background

noise for the live sports broadcast. The recorded signals were passed through a microphone preamp with an Analog-Digital converter interface that allowed the signals to be recorded in a computer. Then, the digital audio signals were recorded using a specific software developed by our research group with an interface that showed the script when necessary.

5.3 Prediction based TTS systems

In section 2.4 we have already analyzed the basics of genre prediction techniques, but in order to apply them in a workable system they must be particularized. For the present thesis we considered a centroid-based predictor, which is a simple but effective implementation.

5.3.1 Centroid based predictor

This classifier works by assuming that the centroid of all the document vectors assigned to the same class (genre in our particular application) is the best representation of that class. The centroid $\vec{C}_z$ are obtained following the Arithmetic Averaging Centroid (AAC) formula [Han and Karypis, 2000], which considers the sums of the weights of the terms present in the n_z documents d_j of genre a_z :

$$\vec{C}_z = \frac{1}{n_z} \sum_{d_j \in a_z} \vec{d}_j \quad (5.1)$$

Then, we need to obtain the similarity measure of the text to be synthesized with the prospective genres. For the proposed system we use a distance measure instead of a similarity one as it will be more useful for the semi-continuous system:

$$dist(\vec{x}_i, \vec{C}_z) = 1 - \frac{\vec{x}_i' \cdot \vec{C}_z}{|\vec{x}_i'| \cdot |\vec{C}_z|} \quad (5.2)$$

Finally, the classification process is reduced to finding the minimum distance between the input text and all the centroids:

$$\hat{c}_* = argmin_{\vec{C}_z} \{dist(\vec{x}_i, \vec{C}_z)\} \quad (5.3)$$

5.3.2 Discrete approach

The discrete system is a direct application of the centroid based predictor, and works by inputting a text to be synthesized to the genre predictor and then using the classification result to select an adequate speaking style synthesis model without any additional information. Naturally this approach is a significant step-up from the traditional TTS systems that are not capable of adapting to the input text. In figure 5.2 we can see a schematic of the system capable of selecting the most adequate speaking style model according to the predicted genre of input text, and then synthesizing it.

Figure 5.2: Schematic of the proposed discrete approach to the expressive TTS system.

As we can see in the figure, the process is clearly divided into the two traditional steps of any TTS system, training and synthesis. In terms of training, we can consider two separate processes, one for the speaker modeling and another one for the CVSM modeling. The speaker modeling follows the already explained process (sections 3.2.2, 4.2.2 and others) of first training the AVM to provide a context-rich background model for the following adaptations that provide the actual expressive and neutral models. The CVSM modeling in this case consists in estimating the SVD matrices T and S that will allow us to project the utterances of the synthesis step, and also the centroids $\vec{C}_z$ that will enable the genre prediction step of the synthesis step.

The synthesis step consists in first projecting the input text to be synthesized as a query to the centroid predictor as follows:

$$\vec{x}'_i = \vec{w}_q^T \cdot \hat{T} \cdot \hat{S}^{-1} \quad (5.4)$$

Where $\vec{w}_q$ is the weighted query vector defined in equation 2.4.2. Finally the genre prediction itself is carried out, and its result is used to automatically select the appropriate expressive model and used to synthesize the input text.

5.3.3 Semi-continuous approach

The second proposed system makes use of semi-continuous genre prediction coupled with discrete style modeling. This means using the distance of the input text vector to the vectors that define the genre centroids as a way of automatically obtaining an expressive strength value that we can use as a model interpolation factor to control the produced expressive strength. This would increase the adaptability capabilities of the system, as not only the synthesized speaking style is obtained from text but also the expressive strength. Due to the nature of our expressive TTS system, we want the expressive strength to take values between 0 and 1 where 0 is the neutral speaking style and 1 is the maximum expressive strength (which could go further in case we wanted to exaggerate the speaking style).

One way of obtaining the desired control factor k_{c^*} is directly using the minimum distance measure. As the cosine distance (as defined in 5.3) ranges between 0 and 2, the simplest approach would be to simply apply the following formula:

$$k_{c^*} = 1 - \frac{\hat{c}_*}{2} \quad (5.5)$$

This simple normalization, while useful, lacks discrimination power. In the situation where the text to be synthesized is very similar to two different genres it is not capable of considering the low confidence of the prediction.

Other possible control factors that take into account not only the distance to the closest genre but also to the second closest or even to every other genre can be used, but all of them have certain conditions in which they lack discrimination power. For that reason we decided to introduce a

frontier based control factor. This approach is based on the concept that after all CVSM-based genre prediction considers an n-dimensional vector space, where a vector that defines the frontier between genres i and j $\vec{f}_{ij}$ can be defined as:

$$\vec{f}_{ij} = \frac{\vec{C}_i + \vec{C}_j}{2} \quad (5.6)$$

Where $\vec{C}_i$ and $\vec{C}_j$ are the centroids of two genres. Then, to obtain the frontier-based interpolation factor, we obtain the frontier vector for the two closest centroids. Finally the proposed interpolation factor becomes the distance between the vector of the input text and the frontier:

$$k_{C^*} = d_f = \text{dist}(\vec{x}'_i, \vec{f}_{12}) = 1 - \frac{\vec{x}'_i \cdot \vec{f}_{12}}{|\vec{x}'_i| \cdot |\vec{f}_{12}|} \quad (5.7)$$

Where $\vec{x}'_i$ is the CVSM projection of the text to be synthesized and $\vec{f}_{12}$ the frontier between the closest and second closest genres.

The final problem to solve is the generalization of the results. The resulting distance measure depends on many factors besides the text itself: the number of dimensions considered for the system, the training data, etc. So it becomes important to create a normalization system that converts the predicted interpolation factor to a value usable in our expressive speech synthesis [Lorenzo-Trueba et al., 2015]. Our studies have shown that our system is capable of handling paralinguistic features control factor between 0.5 and 1.25 for speaking styles synthesis, where 0.5 is already significantly perceived as the target expressiveness, and 1.25 still maintains a good speech quality while providing increased perceived expressive strength (see annex A, sections A.1 and A.2).

The proposed normalization strategy is the use of a Sigmoid function to distribute the values between 0,5 and 1,25 with the particularity that we want the input values equal to the mean distance μ to be associated to 1. The function can be then obtained by solving c in:

$$S_d(\mu) = \frac{0,75}{1 + e^{\frac{-(\mu - \mu + c)}{\sigma/2}}} + 0,5 = 1 \quad (5.8)$$

which, when solved becomes:

$$c = -\left(\frac{\sigma}{2} \cdot \ln 0,5\right) = 0,34657\sigma \quad (5.9)$$

finally, the desired Sigmoid function is:

$$Sf(t) = \frac{0,75}{1 + e^{\frac{-(\mu - \mu + 0,34657\sigma)}{\sigma/2}}} + 0,5 \quad (5.10)$$

where the μ and σ must be obtained with a development set as the mean distance of the development texts to their respective frontiers and its standard deviation respectively.

Figure 5.3: Schematic of the proposed semi-continuous approach to the expressive TTS system.

In figure 5.3 we can see the schematic of the proposed semi-continuous system. It is very similar to the discrete system in figure 5.2 with some critical differences: in the training section, instead of simply adapting into the expressive speaker models, the adaptation functions are obtained so that they can be used in the paralinguistic features control step that is carried out at synthesis time. Also, there are some differences at synthesis time: namely the genre prediction is capable of providing the control factor k_{C^*} described in this section, which is then used in conjunction to the genre prediction to produce the controlled using the system explained in chapter 3.

5.3.4 Perceptual Evaluation

The perceptual evaluations had two goals in mind: first to evaluate the recorded database to verify that it fulfills some standards of quality and also that the recorded styles are actually adequate for the genres that we wanted to convey and to what degree it was conveyed by the professional speaker. Secondly, we wanted to verify the usefulness of the proposed automatic expressive speech synthesis systems in terms of adequacy, speech quality and perceived expressive strength with varying genre prediction error rates to quantify the advantages of having a good genre predictor. This was done by means of three evaluations: a first one that considered the natural voice of the recorded speaker, a second one that considered the discrete version of the system together with the traditional neutral synthetic speech and a final one that evaluated the semi-continuous version of the proposed system together with the neutral synthetic voice too. All three variables in the three evaluations were measured by means of a 5-points MOS scale. This enabled us to compare our proposed systems with the neutral baseline and the natural voice top-line and as such obtain representative results.

The proposed genre prediction system was evaluated by means of cross-fold validation with the training set of SSS split into 90% train and 10% test folds, using the averaged results also as the sigmoid parameters. We also consider a number of genre prediction systems with fixed prediction rates. This is done by manually changing the dimensionality of the CVSM projection, which will directly affect prediction rates. The relationship between prediction rates and CVSM dimensionality for the perceptual evaluation texts can be seen in figure 5.4. There we see how error rates decrease with the number of dimensions until we reach the maximum of n (i.e. training data number of documents). Thus, picking a system with a desired prediction rate means directly choosing the adequate dimensionality. Ultimately, evaluating the different prediction rate systems allows us to analyze how the proposed automatic expressive speech synthesis system would fare when considering different evaluation or training data which could present different prediction rates to our particular implementation.

Evaluations Design

Four speaking styles (news broadcasting, interviews, live sports commentary and political speech) were analyzed in all three evaluations. In order to evaluate the performance of the proposed automatic speech synthesis with different genre prediction error rates. We considered 4 possible systems: 40%, 60%, 80% and 100% prediction rates. Together with the neutral system this re-

Figure 5.4: Schematic of the proposed semi-continuous approach to the expressive TTS system.

sulted in a total of 20 systems to evaluate (4 rates, 5 genres). Following the Latin-square [Gao, 2005] approach this meant that we needed 20 different utterances to be synthesized for all the systems to be presented to the listeners in a random order without repetitions. In the end we decided for considering 20 utterances per evaluation and not multiples of it so as to keep the evaluation itself from becoming too long, lasting around 20 minutes in this implementation. For the natural voice evaluation we considered the same approach of 20 utterances evaluated per listener in order to obtain more significant results faster even if there were only 4 systems to evaluate.

The test itself was carried out by means of a web interface, where the evaluator was presented with a button to play the audio sample and the transcription of the uttered texts. The samples could be played as many times as desired. Then, the listener was asked to rate the utterances in the traditional 5 point MOS evaluation in terms of adequacy of the utterance to the text (from not adequate to very adequate), speech quality (from very bad to very good) and perceived expressive strength when comparing to a neutral version (very low to very high). The speaker models were trained following the usual procedure (see section 4.3) using the whole train section of SSS database. For the semi-continuous version the interpolation was carried out at a model level [Lorenzo-Trueba et al., 2012c].

In terms of the statistical significance of the results, we applied the Wilcoxon Signed-Rank Test for a 95% confidence ratio in order to obtain the error margins. A minimum of 20 subjects per evaluation was fixed in order to guarantee the full coverage of the Latin square matrix, totaling

	Interview	News	Speech	Sports
Interview	95	4	1	0
News	1	96	3	0
Speech	0	1	99	0
Sports	0	1	0	99

Table 5.3: Train texts genre predictor confusion matrix in percentage. Speech stands for political speech and sports for live sports broadcasts. Rows show the predicted style and columns the actual style.

	Interview	News	Speech	Sports
Interview	90	5	5	0
News	0	100	0	0
Speech	0	5	95	0
Sports	0	0	5	95

Table 5.4: Evaluation texts genre predictor confusion matrix in percentage. Speech stands for political speech and sports for live sports broadcasts. Rows show the predicted style and columns the actual style.

at least 20 samples per speaking style and system¹.

Genre Prediction System Results

The results of the CVSM-based genre predictor trained with only the train set and applied to the test show a genre prediction rate of 90% when averaged across the 10-folds when trained using 333 dimensions.

We can see the confusion matrices for both train and evaluation texts in tables 5.3 and 5.4. There we can see how the results are very good in both cases, reaching a 95% prediction rate for the evaluation texts, higher than the three forced error rates of the perceptual evaluation.

Database Evaluation Results

20 listeners participated in the database perceptual evaluation, for a total of 500 samples, 100 samples per style. The results can be seen in table 5.5, and prove the high quality of the database in the three considered factors with low error margins. Namely, the results show an average across all four styles of 4.2 adequacy, 4.1 expressive intensity and 4.5 speech quality.

¹ Samples can be listened from <http://lorien.die.upm.es/~jaime.lorenzo/thesis/system.htm>

	Interview	News	Speech	Sports
Adequacy	4.20	4.19	3.93	4.26
Intensity	4.19	4.07	3.82	4.31
Quality	4.52	4.50	4.45	4.45

Table 5.5: Results in MOS scale for the perceptual evaluation of the SSS database with error margins for a 95% confidence rate. Intensity stands for expressive intensity and quality for speech quality. Also, speech stands for political speech and sports for live sports broadcasts.

For the results of the proposed systems perceptual evaluations, and in order to make their comparison between the baseline and top-line systems easier, figure 5.5 to 5.10 include the results for the natural speech evaluation.

Discrete System Results

In figures 5.5 to 5.7 we can see the results of the perceptual evaluation for the three discrete systems (40D, 60D, 80D and S which refer to the 40%, 60%, 80% and 100% prediction rate systems respectively), the traditional neutral system N and the natural voice NAT. The four speaking styles are live sports broadcasts (SP), political speech (PO), interviews (IN) and news broadcasts (NW). At first 20 people took part on the evaluation, but aiming to increase the statistical significance of the results 10 more people participated, for a total of 600 samples. Increasing to 30 listeners practically maintained the error margins because of the variance in the results, so we did not increase the numbers any further. This high variance is mainly due to the influence of prediction errors, which means that for the lower prediction rate systems variance can be expected to be much higher than those of higher prediction rate systems or natural speech. As such, error margin will also be higher for the 40% or 60%, making it harder to draw significant results from them.

In terms of adequacy results (figure 5.5) we can see how for the live sports speaking style we are able to significantly improve the results over the traditional neutral system and also how there is an increasing adequacy together with the increase in prediction rates to a maximum of 4.27 for the 80% rate system which is practically the same than the 4.26 attained for the natural voice. For other speaking styles besides interviews we can see that there are consistently increasing values in adequacy for higher prediction rates until the 80%-100% system practically reaches the natural voice adequacy.

The perceived expressive intensity results (figure 5.6) also show significant differences in the

Figure 5.5: Results in MOS scale of the adequacy evaluation for the discrete system.

Figure 5.6: Results in MOS scale of the expressive intensity evaluation for the discrete system.

live sports broadcasting style, where there is a sharp increase in perceived expressive strength from the 40% to the 60% prediction rate system, and from there the values remain similar to the 4.31 attained in the natural voice evaluation. In the other speaking styles, once again excluding interviews, we can see not statistically significant but notable (close to 1 point for political speech and 0.5 for news broadcasting) differences from the neutral speech to the expressive systems but without too much differences between the expressive system themselves.

Regarding speech quality (figure 5.7) we can see how in general the natural voice presents a significantly higher value (4.50 in average) for all speaking styles, which is significantly better than the 3.5 average quality of the neutral synthetic speech. In this case it is only in sports broadcasts style where we can see also significant differences between the proposed expressive

Figure 5.7: Results in MOS scale of the quality evaluation for the discrete system.

system and the traditional neutral speech for the 80% and 100% prediction rates system. For the other speaking styles there is no relevant difference.

Semi-continuous System Results

The semi-continuous evaluation results, show in figures 5.8 to 5.10, follow the same pattern as the discrete evaluation, where the four proposed systems are represented by 40SC, 60SC, 80SC and S, the traditional neutral system by N and the natural voice by NAT. The same abbreviations for the speaking styles are used. It is important to mention that the 100% system undergoes no interpolation as there is no way to obtain the interpolation value because it is not the output of the genre predictor but an oracle situation. In this evaluation, for the same reasons as in the discrete system evaluation, also a total of 30 listeners participated, for a total of 600 samples.

For the proposed semi-continuous systems adequacy (figure 5.8), we are able to achieve significant differences in both live sports broadcasts and political speech, reaching almost significant differences with the proposed system at 80% prediction rate for sports when compared to natural speech with an adequacy value of 4.5 compared to 4.3. In the case of political speech the results are similar from the 60% prediction rate system onwards maxing at 3.8 for the 80% system, which is significantly better than the 2.6 adequacy of the neutral speech system and comparable to the 3.9 adequacy reached by natural speech. Interview once again shows worse results for all the synthetic versions when compared to the natural speech and news almost significantly improves the neutral system with a 3.9 adequacy for the 60% to 100% systems when compared

Figure 5.8: Results in MOS scale of the adequacy evaluation for the semi-continuous system.

to the 3.3 adequacy of the neutral voice, but is, in any case, not as high as the 4.2 of the natural voice.

Figure 5.9: Results in MOS scale of the expressive intensity evaluation for the semi-continuous system.

In the perceived expressive intensity (figure 5.9) we break the trend of the adequacy results, as we obtain statistically significant differences for live sports broadcasts, political speech and news broadcasts and not statistically significant differences only for interviews. In this evaluation we can see, outside from interview results, how there is a trend of higher perceived emotional strength for better genre prediction rates, capping at the natural voice results. In the case of live sports broadcasts there are almost significant differences between the 60% and the 80% prediction rate systems, reaching values as high as 4.4 compared to the 1.9 perceived intensity of the neutral synthetic speech and the 4.3 of natural speech. It is also notable how, even if not significantly, there is a reduction in perceived expressive strength from the 80% system to the 100% system due to it not being interpolated.

Figure 5.10: Results in MOS scale of the quality evaluation for the semi-continuous system.

The results of speech quality for the semi-continuous systems (figure 5.10) are similar to those of the discrete system, where only live sports broadcasts achieve quality similar to that of natural speech, improving the neutral synthetic speech from 3.7 to 4.3 in the 80% system. In the other three speaking styles natural speech is significantly better than the achieved synthetic speech quality as usual, although for political speech we are close to reducing the breach.

5.4 Continuous system

The ideal expressive TTS system should be able to handle any kind and any number of expressiveness without requiring costly labeling or manipulations every time we want to add any new one. With this purpose in mind the concept of a continuous system (i.e. a system in which ex-

pressiveness is treated as a complete space instead of as discrete entities) fits perfectly. That is, in real life expressiveness present overlap between them because there are not clear frontiers, so being able to take into account those overlaps in the shape of a dynamic synthesizer would be ideal.

For that objective we propose a continuous system based both in CVSM and MR-HMM, continuous vector space modeling to model the expressive training texts as a continuous space and multi-regression to make use of that continuous space as auxiliary features for the HMM modeling. This results in a system capable of characterizing input texts as a vector, which at the same time ends up producing the adequate acoustic model associated to the input text. As such, as long as sufficient variety in the training data is available, we would be able to model a continuous lexical space into a continuous acoustic space.

5.4.1 Proposed System

Figure 5.11 shows the proposed flowchart for the continuous, MRHMM-based system. There we can see many fundamental differences with the discrete and semi-continuous approach, both at training time and at synthesis time.

Figure 5.11: Schematic of the proposed continuous approach to the expressive TTS system.

Most notably, there is no traditional adaptation process involved in the system, but a MR-HMM adaptation that takes the CVSM projections of each training text file as the control vector $\vec{x}'$ for a training process that outputs a new model that combines both neutral and expressive information. Also we can see how there is no need for a centroid estimation, as there is no genre prediction carried out at synthesis time. Instead, only the CSVM matrices are kept as the output so that at synthesis time the synthesis text can be projected as explained in equation 5.4 to obtain $\vec{x}'_i$, which will be used directly at the speech generation process. This process is then capable of producing an expressive speech output without relying in any genre prediction system, only in the CVSM projections of the synthesis texts.

5.4.2 Perceptual Evaluation

The evaluated implementations of the proposed system considered a 4-dimensional space for the CVSM, one per considered speaking style. This implies a reduction in discrimination capabilities (the centroid-based prediction rates for 4 dimensions goes down to 76%), but it is necessary because we cannot expect MR-HMM to work correctly with more than 300 auxiliary parameters, which was confirmed by informal listening tests. The value is obtained by grouping all the training texts associated to the same speaking style to the same document, which automatically fixed the result to 4 dimensions, but with a higher amount of speaking styles the dimensionality would increase accordingly.

We also considered two approaches to the MR-HMM modeling: a first one that directly utilized the CVSM-projected vectors $\vec{x}'$ and a second one that normalized each component by the maximum of their respective component $\vec{x}' = \{v_1/v_{MAX1}, \dots, v_4/v_{MAX4}\}$ where $v_{MAXi} = \max(v_i) \forall v_i$ in the training data. The second approach was considered in case reducing the dispersion of the control vector values helped the MR-HMM re-estimation process. The same normalization was applied to the synthesis control vectors.

The evaluation itself was practically the same as the continuous and semi-continuous evaluations with the exception of the number of utterances, which was set to 8 because some of the test sentences acoustic parameters resulting from the MR-HMM modeling were not synthesizable (i.e. frames with infinite MCEP). The rest was the same, Latin-square approach to remove learning bias and MOS for the three evaluated variables. The vocoder configuration was also the same. The evaluation of the normalized system and the non-normalized systems were carried out independently, for a total of 32 listeners, 16 per evaluation, and 512 total utterances.

Synth. Rate	C_Normalized	Continuous
Interview	48%	43%
News	81%	67%
Speech	81%	90%
Sports	52%	38%
Average	65%	60%

Table 5.6: Percentage of synthesizable test sentences for each system-style pair. C_Normalized represents the normalized implementation of the continuous system and Continuous the non-normalized one.

Evaluation Results

Figures 5.12 to 5.14 show the results for both continuous system evaluations (C_Normalized for the normalized version and Continuous for the non-normalized one), together with the best results from the discrete and semi-continuous system evaluations (80D and 80SC), neutral speech (N), traditional emotional system (S) and natural voice (NAT). It is important to emphasize that this evaluation considered only utterances that could be synthesized, which represented a 65% for the normalized version and 60% for the non-normalized one, details for each system-style interaction can be seen in table 5.6.

In the adequacy results (figure 5.12) we can see how the system provides significant increases when compared to traditional neutral systems, although not as big as those of the semi-continuous system. In the case of the normalized system, the average adequacy is 3.47 and 3.60 for the non-normalized version, not significantly worse than the 3.78 of the traditional expressive synthesis. Interviews, again, due to its conversational nature was not modeled adequately.

Perceived expressive intensity results (figure 5.13) show a similar image to adequacy. Comparable results to the other systems while being significantly better than using a neutral system. Both proposed continuous systems show a 3.51 average perceived expressive intensity, which is once again comparable to the 3.65 of the non-predictive speech synthesis, although far from the 4.07 of natural speech and much better than the 2.51 of neutral speech.

Finally speech quality (figure 5.14) does show some bad results, an average quality of 2.93 for the basic system and 2.73 for the normalized system, which is significantly worse than all other systems. This is mainly because the continuous system introduces a large amount of artifacts. This effect was somewhat expected due to the inherent semi-continuous nature of the SSS database: the professional speaker was asked to interpret the four speaking styles showing as little varia-

Figure 5.12: Results in MOS scale of the adequacy evaluation for the continuous system.

Figure 5.13: Results in MOS scale of the expressive intensity evaluation for the continuous system.

tion as possible so that they were clear representatives of their paralinguistic characteristics, so modeling them in a continuous fashion is not possible without additional data. More varied data or a more naturally continuous task is expected to fare better for the continuous system.

5.5 Systems Comparison

From table 5.7 to 5.9 we can see a comparison of the results of the 6 considered systems. The traditional neutral read speech TTS system, the non-normalized continuous system, the best discrete system (80% prediction rate), the best semi-continuous system (80% prediction rate),

Figure 5.14: Results in MOS scale of the quality evaluation for the continuous system.

Adequacy	Neutral	Discrete	Semi-continuous	Continuous	Synthetic	Natural
Sports	2.07	4.27	4.50	3.56	4.15	4.26
Speech	2.57	3.83	3.83	3.59	3.78	3.93
News	3.33	3.90	3.87	3.76	3.85	4.19
Average	2.66	4.00	4.07	3.64	3.93	4.13

Table 5.7: Adequacy results in MOS scale for all the compared expressive TTS systems.

the traditional oracle expressive TTS and NAT for the results of the natural voice. Interview style was not compared because it was already determined that none of the considered systems is capable of correctly modeling it.

In terms of adequacy (table 5.7), we can see how the three proposed systems provide significantly better results than traditional neutral speech and comparable to natural speech, except in the case of interviews. This proves that all the proposed systems are capable of significantly increasing the expressive capabilities of traditional TTS systems in an automatic way that relies only in lexical information instead of on expert information as traditional expressive TTS systems do, which is a very good result.

Intensity	Neutral	Discrete	Semi-continuous	Continuous	Synthetic	Natural
Sports	1.97	4.23	4.43	3.56	4.11	4.31
Speech	2.57	3.63	3.80	3.44	3.56	3.82
News	2.93	3.63	3.83	3.59	3.70	3.97
Average	2.49	3.83	4.02	3.53	3.79	4.03

Table 5.8: Perceived expressive intensity results in MOS scale for all the compared expressive TTS systems.

Quality	Neutral	Discrete	Semi-continuous	Continuous	Synthetic	Natural
Sports	3.70	4.33	4.27	2.82	4.33	4.45
Speech	4.00	3.77	3.83	3.09	4.10	4.45
News	3.93	3.73	4.00	3.18	3.90	4.50
Average	3.88	3.94	4.03	3.03	4.11	4.47

Table 5.9: Speech quality results in MOS scale for all the compared TTS systems.

For perceived expressive intensity results (table 5.8) we can gather the same conclusions as with adequacy results. The proposed systems are capable of producing perceived expressive intensities comparable to those of natural speech excepting for interview styles, although in this situation we are capable of improving even traditional expressive speech synthesis thanks to the expressive strength control capabilities of the systems. As such, it becomes clear that an additional advantage of the proposed systems is that it is capable not only of automatically generating expressive speech synthesis but also of emphasizing the paralinguistic features to enhance the expressive capabilities, overcoming the inherent smoothing of HMM-based synthesis.

Finally, for speech quality (table 5.9) we can see how we are still not capable of reaching natural voice level results. In the case of the discrete and semi-continuous systems speech quality results are very similar to those achieved by more traditional and polished systems such as neutral read speech, which means that the proposed systems are capable of providing the increased expressive capabilities without incurring in a speech quality cost as it happened in the proposed transplantation system. On the other hand we can see a sharp decrease in speech quality for the continuous system, which we assume to be due to the non-continuous nature of the data used to train the system, that fits better a discrete or semi-continuous approach.

In conclusion we can see how the designed and evaluated expressive TTS systems are more than capable of improving the expressive capabilities of traditional TTS systems, be it because the system is capable of automatically adapting to the text to be synthesized in order to provide the most adequate speaking style, or because the system is capable of controlling the produced

expressive strength to match the representativeness of the target text when compared to prototypical genres. We can also see how the proposed continuous system is capable of achieving this without any kind of lexical labeling or genre prediction. We expect the continuous system to be much more versatile than the discrete or semi-continuous approach in a situation where the data correctly reflects the continuous nature of the problem, which remains to be studied.

5.6 Expressive TTS System Conclusions

We have recorded the first single speaker Spanish speaking styles database, which was designed with the task of speaking styles synthesis in mind. Four speaking styles were recorded (live sports broadcasts, news broadcasts, political speech and interviews) in order to cover different degrees of improvisation and formality in terms of media styles. This database consists in a total of 834 utterances split into a 750 train block and 84 test block (21 test utterances per style), and approximately 1h of material for each speaking style.

We have also developed and evaluated three different approaches to the expressive speech synthesis problem that rely on automatic speaking style selection based on genre prediction from the text to be synthesized, and one that only relies in lexical information. All three systems rely on continuous vector space modeling, which not only provides a highly reliable genre prediction, with average prediction rates of 90% for the recorded database, but also provides a way to represent lexical information that can be used to modify speaker models thanks to MR-HMM techniques. The first considered approach, the discrete system, makes use of genre prediction directly to select the most adequate speaking style model and synthesize accordingly. The semi-continuous system makes use of the genre prediction coupled with a measure distance of the text to be synthesized to the frontier between the predicted genre and the second candidate to provide a control factor for the paralinguistic features control system also proposed in this thesis. This allows the semi-continuous system to not only automatically select the most adequate speaking style to the text, but also to automatically control the synthesized expressive strength according to how similar the synthesized text is to the chosen expressiveness model. Finally, the continuous system is capable of synthesizing the most adequate expressiveness without relying on any genre prediction, just by making use of the CVSM projection of the input text as a control factor of the speaker model, which vastly increases the versatility of the system as we remove the need for labeling the training data genre.

Finally we have carried out a number of different perceptual evaluations. A first one to validate the introduced database, which proved that the recorded styles are highly adequate to the genre of the text they are representing (average of 4.1) and also of very high quality (average of 4.5) and intensity (average of 4.1). A second one to measure the effects of the proposed discrete system, where it was shown that it is for an 80% prediction rate is capable of consistently improving the traditional neutral speech synthesis systems and significantly for the live sports broadcast style and coming close to the natural voice in all the styles excepting interviews. For the best discrete system, excluding interviews, we obtained an average of 4.0 in adequacy, 3.83 in intensity and 3.9 in speech quality. The third evaluation focused on the semi-continuous system, which was proven to also consistently improve the neutral synthesizer in general and also providing more significant results not only on sports broadcasts but also on political speech when compared to the discrete system. The semi-continuous system also showed a better capability of conveying expressive strength than the discrete system, coming closer to the results of the natural voice evaluation (average scores of 4.1 in adequacy, 4 in intensity and 4 in quality). Finally we also evaluated the continuous system, which proved to be capable of conveying the correct expressiveness (average adequacy of 3.6) with an expressive strength not statistically different to the semi-continuous system (average of 3.6) although at a significant drop in speech quality mainly due to the semi-continuous nature of the data (average quality of 2.9).

All in all we have introduced a different approach to expressive speech synthesis where the system automatically adjusts the produced speaking style according to the text to be synthesized without requiring any output from the user besides previously training it to recognize the text genres to be considered. The systems have shown that they are capable of significantly improving the traditional neutral speech synthesis systems in the task, and also capable of providing similar adequacy and perceived expressive strength rates than those of natural voice. Also it has been shown that for SSS the semi-continuous system is the most adequate even if the other systems provide comparable adequacy and expressive strength rates. Most notably in the case of the sport broadcasts speaking style we have been able to improve the results of natural voice in terms of adequacy and produced expressive intensity.

Chapter 6

Conclusions

In general we can say that all of the objectives have been achieved successfully. The paralinguistic features control system is capable of enhancing or attenuating the expressiveness of expressive models including emotional speech or speaking styles to a point that was not achievable with traditional systems based on interpolation. The features control system also prepared the way for the paralinguistic information transplantation system, as it allowed us to develop a system capable of controlling not only expressive information but also speaker identity information in order to imbue expressiveness into target speakers for whom only neutral data was available.

Finally we have also developed a complete and autonomous system capable of making use of the already presented systems just by relying on the input text information instead of requiring user input. That is, the proposed expressive TTS system is capable of automatically predicting the most adequate expressiveness to synthesize the desired text so as to maximize naturalness. This has been done following different approaches, with increasing degrees of freedom: the first approach (discrete) is only capable of selecting the most adequate speech model, the second approach (semi-continuous) is capable of selecting the speech model and of modulating the expressive intensity according to the text. Finally we have also developed a continuous system that models all expressiveness in a single model by making use of lexical features instead of relying on genre labels. All three systems have proven to be capable of adequately synthesizing speaking styles with expressive intensities comparable to natural voice and quality similar to traditional HMM-based synthesis, although the continuous system is not capable of providing adequate speech quality because our data is intrinsically semi-continuous and is not correctly modeled in a continuous fashion.

Additionally we have recorded the first single speaker speaking styles database in Spanish, which is what allowed us to correctly work in speaking styles synthesis. We have also developed a cross-lingual emotion transplantation system that improves the limited cross-lingual works that have been carried out. The cross-lingual system development process also helped us develop a model for predicting possible artifacts in adapted models by making use of context-to-context distances. This prediction model, while not sufficient by itself to reliably predict artifacts, showed significant evidence that the higher the distance between contexts the higher the probability of an artifact appearing.

6.1 Future Work

For future work we want to consider a broader array of expressiveness, in order to find a problem where the continuous modeling fits naturally: a more complete speaking styles collection, or even considering sub-spaces of speaking styles. Our emotional databases, unfortunately, while being continuous in nature do not have such a nature translated into words, so they do not seem to be a good fit for the continuous system either. So recording new databases will be mandatory.

Another field that we want to work on is on bringing our systems into the DNNs and RNNs world. This task will prove challenging as there is still not many researches underway on DNN-based expressive speech synthesis. Cross-lingual transplantation is also another field where there is still much work to be done, particularly for applying information easily learned from majority languages into minority languages. This is currently state of the art in DNN-based synthesis, but once again expressiveness has not been a priority so we should focus on that too. Concretely, trying to model long-term prosody by means of a RNN to then codify it by means of an auto-encoder could help overcome the limitations of our proposed KLD-based state mapping approach. Adapting the paralinguistic features control and transplantation systems are expected to be more of a challenge because implementing factorization in neural networks is not trivial.

Finally, carrying out evaluations in real life systems such as car navigation systems or robotic assistants in scenarios not as constrained as the ones we evaluated would provide much needed information on how research systems fare in the real world, which would undoubtedly give hints on where more to focus our efforts. This particularly means focusing on Lombard effect, as it is an omnipresent effect that humans unconsciously use to improve conversation intelligibility under noisy conditions. If we are able to correctly apply this effect into real life systems where

noise is unavoidable (e.g. train station dialog systems, car navigation systems in some likely circumstances), we can expect to significantly increase the intelligibility and useability of speech technologies in everyday situations. We have high hopes on this approach based on the good results achieved in the sports broadcasts style, which relied in its recording on a Lombard-like effect (stadium noise to provide ambient for the speaker), which helped us even improve natural voice results.

Chapter 7

Contributions

7.1 Refereed Journal Papers

- **J. Lorenzo-Trueba**, R. Barra-Chicote, R. San-Segundo; J. Ferreiros, J. Yamagishi, J.M. Montero, "*Emotion Transplantation through Adaptation in HMM-based Speech Synthesis*". Computer Speech and Language, vol 34, pp 292-307, 2015.
- S. Lebai Lutfi, F. Fernández-Martínez, **J. Lorenzo-Trueba**, R. Barra-Chicote and J. M. Montero "*I Feel You: The Design and Evaluation of a Domotic Affect-Sensitive Spoken Conversational Agent*", Sensors 2013, 13(8), 10519-10538.

7.2 Book Chapters

- **J. Lorenzo-Trueba**, R. Barra-Chicote, J. Yamagishi, and J. M. Montero, "*Towards Cross-lingual Emotion Transplantation*". in Advances in Speech and Language Technologies for Iberian Languages, Lecture Notes in Computer Science Volume 8854, 2014, pp 199-208.

7.3 Refereed Conference and Workshop Papers

- **J. Lorenzo-Trueba**, J. D. Echeverry-Correa, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, A. Gallardo-Antolín, J. Yamagishi, S. King and J. M. Montero, "*Development of a*

Genre-Dependent TTS System with Cross-Speaker Speaking-Style Transplantation". Proceedings of the 2nd International Workshop on Speech, Language and Audio in Multimedia (SLAM2014), Penang-Malaysia 11-12 September 2014.

- S. Lutfi, F. Fernández-Martínez, **J. Lorenzo-Trueba**, R. Barra-Chicote, J.M. Montero "*NEMOHIFI: An Affective HiFi Agent*". ICMI 13, 15th ACM International Conference on Multimodal Interaction, Sydney Australia 2013.
- **J. Lorenzo-Trueba**, R. Barra-Chicote, J. Yamagishi, O. Watts, J. M. Montero "*Towards Speaking Style Transplantation in Speech Synthesis*", Proceedings SSW8 2013 - 8th ISCA Speech Synthesis Workshop, August 31st - September 2nd, 2013, Barcelona, Spain.
- **J. Lorenzo-Trueba**, R. Barra-Chicote, J. Yamagishi, O. Watts and J.M. Montero "*Evaluation of a Transplantation Algorithm for Expressive Speech Synthesis*", Workshops en Tecnologías Accesibles parte del IV Congreso Español de Informática CEDI2013 17-20 September 2013, Madrid (Spain).
- **J. Lorenzo-Trueba**, O. Watts, R. Barra-Chicote, J. Yamagishi, S. King, and J.M. Montero. "*Simple4all proposals for the albayzin evaluations in speech synthesis.*", IberSPEECH 2012 - VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, pp. 653-663.
- V. López-Ludeña, R. San-Segundo, J. M. Montero, R. Barra-chicote, J. Lorenzo-Trueba, "*Architecture for Text Normalization using Statistical Machine Translation techniques*", IberSPEECH 2012 - VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, pp.204-213.
- **J. Lorenzo-Trueba**, B. Martínez-González, R. Barra-Chicote, V. López-Ludeña, J. Ferreiros, J. Yamagishi and J.M. Montero, "*Towards an Unsupervised Speaking Style Voice Building Framework: Multi-Style Speaker Diarization*", InterSpeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13, 2012.
- **J. Lorenzo-Trueba**, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi and J.M. Montero, "*Towards Glottal Source Controllability in Expressive Speech Synthesis*", InterSpeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13, 2012.

7.4 Other Contributions

- Winner of the Albayzin 2012 Speech Synthesis Evaluation
- Design and record the first single speaker Spanish Speaking Styles database. Soon to be released for research purposes.
- Implementation of cascade adaptation in HTS for both VCTK and VCTK2 frameworks.
- Discovery and fixes for the state mapping functions that were present in the 2.1.1 version of HTS, included in HTS version 2.3.

Bibliography

- Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.
- Abe, M. (1997). Speaking styles: Statistical analysis and synthesis by a text-to-speech system. In Santen, J., Olive, J., Sproat, R., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 495–510. Springer New York.
- Adell, J., Bonafonte, A., and Escudero-Mancebo, D. (2010). Modelling filled pauses prosody to synthesise disfluent speech. *Proc. ISCA Speech Prosody, Chicago, USA*.
- Adell, J., Escudero, D., and Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54(3):459–476.
- Aihara, R., Takashima, R., Takiguchi, T., and Ariki, Y. (2012). Gmm-based emotional voice conversion using spectrum and prosody features. *American Journal of Signal Processing*, 2(5):134–138.
- Anastasakos, T., McDonough, J., and Makhoul, J. (1997). Speaker adaptive training: A maximum likelihood approach to speaker normalization. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1043–1046. IEEE.
- Andersson, S., Georgila, K., Traum, D., Aylett, M., and Clark, R. A. (2010). Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. *Speech Prosody*.
- Andersson, S., Yamagishi, J., and Clark, R. A. (2012). Synthesis and evaluation of conversational characteristics in hmm-based speech synthesis. *Speech Communication*, 54(2):175–188.

- Andreas, J. and Ghahramani, Z. (2013). A generative model of vector space semantics. *ACL 2013*, page 91.
- Argente, J. A. (1992). From speech to speaking styles. *Speech communication*, 11(4):325–335.
- Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.
- Balakrishnan, N., Voinov, V., and Nikulin, M. S. (2013). *Chi-squared Goodness of Fit Tests with Applications*. Academic Press.
- Barra-Chicote, R. (2011). *Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis*. PhD thesis, ETSIT-UPM.
- Barra-Chicote, R., Montero, J. M., Macias-Guarasa, J., Lufti, S., Lucas, J. M., Fernandez, F., D’haro, L. F., San-Segundo, R., Ferreiros, J., Cordoba, R., and Pardo, J. M. (2008). Spanish expressive voices: Corpus for emotion research in spanish. *Proc. of LREC*.
- Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., and Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*, 52(5):394–404.
- Benjamin Picart, S. B. and Drugman, T. (2013). Hmm-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation. In *8th ISCA Speech Synthesis Workshop*.
- Berry, M. W., Dumais, S. T., and O’Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.
- Bickmore, T., Bukhari, L., Vardoulakis, L. P., Paasche-Orlow, M., and Shanahan, C. (2012). Hospital buddy: A persistent emotional support companion agent for hospital patients. In *Intelligent Virtual Agents*, pages 492–495. Springer.
- Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1229. IEEE.
- Black, A. W. and Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proc. Eurospeech ’97*, pages 601–604, Rhodes, Greece.

- Blunsom, P., Grefenstette, E., Hermann, K. M., et al. (2013). Not not bad is not bad: A distributional account of negation. In *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*. Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality.
- Bulut, M., Narayanan, S. S., and Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. In *INTERSPEECH*.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.
- Campbell Jr, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462.
- Campillo, F. and Banga, E. (2011). Multiple f0 contour parallel viterbi search for unit selection speech synthesis. *Electronics letters*, 47(16):937–938.
- Charfuelan, M. and Steiner, I. (2013). Expressive speech synthesis in mary tts using audiobook data and emotionml. In *Proceedings of Interspeech*.
- Chen, L., Gales, M., Braunschweiler, N., Akamine, M., and Knill, K. (2013a). Integrated automatic expression prediction and speech synthesis from text. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7977–7981.
- Chen, L., Gales, M., Wan, V., Latorre, J., and Akamine, M. (2012). Exploring rich expressive information from audiobook data using cluster adaptive training. In *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13*.
- Chen, L.-H., Ling, Z.-H., Jiang, Y., Song, Y., Xia, X.-J., Zu, Y.-Q., Yan, R.-Q., and Dai, L.-R. (2013b). The ustc system for blizzard challenge 2013. *Blizzard Challenge 2013*.
- Chesta, C., Siohan, O., and Lee, C.-H. (1999). Maximum a posteriori linear regression for hidden markov model adaptation. In *Eurospeech*.
- Clark, R., Richmond, K., and King, S. (2007). Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, 49(4):317–330.
- Cosma, G. and Joy, M. (2012). An approach to source-code plagiarism detection and investigation using latent semantic analysis. *Computers, IEEE Transactions on*, 61(3):379–394.

- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.
- Dempster, A. P., Laird, N. M., et al. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.
- Desai, S., Black, A. W., Yegnanarayana, B., and Prahallad, K. (2010). Spectral mapping using artificial neural networks for voice conversion. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):954–964.
- D’haro Enríquez, L. F., Glembek, O., Plchot, O., Matějka, P., Soufifar, M., Córdoba Herralde, R. d., and Černocký, J. (2012). Phonotactic language recognition using i-vectors and phoneme posterigram counts.
- Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1):33–60.
- Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.
- Echeverry-Correa, J. (2015). *Contributions to speech analytics based on speech recognition and topic identification*. PhD thesis, ETSIT-UPM.
- Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*.
- Ekman, P. (1999). *Facial expressions*. Wiley Online Library.
- Ekman, P., Friesen, W. V., and Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. Pergamon Press.
- El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.
- Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, recording and verification of a danish emotional speech database. In *Eurospeech*.
- Erro, D., Alonso, A., Serrano, L., Navas, E., and Hernáez, I. (2013a). Towards physically interpretable parametric voice conversion functions. In *Advances in Nonlinear Speech Processing*, pages 75–82. Springer.

- Erro, D., Moreno, A., and Bonafonte, A. (2010a). Voice conversion based on weighted frequency warping. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):922–931.
- Erro, D., Navas, E., and Hernáez, I. (2012). Iterative mmse estimation of vocal tract length normalization factors for voice transformation. In *Proc. Interspeech 2012*.
- Erro, D., Navas, E., and Hernaez, I. (2013b). Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(3):556–566.
- Erro, D., Navas, E., Hernandez, I., and Saratxaga, I. (2010b). Emotion conversion based on prosodic unit selection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):974–983.
- Eskénazi, M. (1993). Trends in speaking styles research. In *Third European Conference on Speech Communication and Technology*.
- Eyben, F., Buchholz, S., Braunschweiler, N., Latorre, J., Wan, V., Gales, M. J., and Knill, K. (2012). Unsupervised clustering of emotion and voice styles for expressive tts. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4009–4012. IEEE.
- Fan, Y., Qian, Y., Soong, F. K., and He, L. (2015). Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4475–4479. IEEE.
- Fan, Y., Qian, Y., Xie, F.-L., and Soong, F. K. (2014). Tts synthesis with bidirectional lstm based recurrent neural networks. In *Interspeech*, pages 1964–1968.
- Fujinaga, K., Nakai, M., Shimodaira, H., and Sagayama, S. (2001). Multiple-regression hidden markov model. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 513–516. IEEE.
- Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 137–140. Ieee.
- Gadallah, M. E., Matar, M., and Algezawi, A. (1999). Speech based automatic lie detection. In *Radio Science Conference, 1999. NRSC'99. Proceedings of the Sixteenth National*, pages C33–1. IEEE.

- Gales, M. J. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech and language*, 12(2).
- Gales, M. J. (2000). Cluster adaptive training of hidden markov models. *Speech and Audio Processing, IEEE Transactions on*, 8(4):417–428.
- Gao, L. (2005). *Latin Squares in Experimental Design*. Michigan State University.
- Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., de-la Mota, C., González, C., Vivaracho, C., Rustullet, S., Larrea, O., et al. (2013). Glissando: a corpus for multi-disciplinary prosodic studies in spanish and catalan. *Language Resources and Evaluation*, 47(4):945–971.
- Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and Audio Processing, IEEE Transactions on*, 2(2):291–298.
- Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.
- Goyal, K., Jauhar, S. K., Li, H., Sachan, M., Srivastava, S., and Hovy, E. (2013). A structured distributional semantic model: Integrating structure with semantics. *ACL 2013*, page 20.
- Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE.
- Han, E.-H. S. and Karypis, G. (2000). *Centroid-based document classification: Analysis and experimental results*. Springer.
- Hecker, M. H. and Williams, C. E. (2005). Choice of reference conditions for speech preference tests. *The Journal of the Acoustical Society of America*, 39(5A):946–952.
- Helander, E., Silén, H., Virtanen, T., and Gabbouj, M. (2012). Voice conversion using dynamic kernel partial least squares regression. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):806–817.
- Henry, E., Hofrichter, J., et al. (2010). Singular value decomposition: application to analysis of experimental data. *Essential Numerical Computer Methods*, 210:81–138.

- Hernández-Figueroa, Z., Carreras-Riudavets, F. J., and Rodríguez-Rodríguez, G. (2013). Automatic syllabification for spanish using lemmatization and derivation to solve the prefix prominence issue. *Expert Systems with Applications*, 40(17):7122–7131.
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.
- Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.
- Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., and Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. In *LREC*.
- Hsu, C.-Y. and Chen, C.-P. (2012). Speaker-dependent model interpolation for statistical emotional speech synthesis. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):1–10.
- Hu, B., Rakthanmanon, T., Hao, Y., Evans, S., Lonardi, S., and Keogh, E. (2014). Using the minimum description length to discover the intrinsic cardinality and dimensionality of time series. *Data Mining and Knowledge Discovery*, pages 1–42.
- Hu, Q., Wu, Z., Richmond, K., Yamagishi, J., Stylianou, Y., Maia, R., King, S., and Akamine, M. (2015). Sinusoidal speech synthesis using deep neural networks. *Signal Processing Letters, IEEE*.
- Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 373–376. IEEE.
- Iberspeech2012 (2012). 2012 albayzin evaluations: speech-synthesis results. In *Proc. Iber-speech*.
- Imai, S., Sumita, K., and Furuichi, C. (1983). Mel log spectrum approximation (mlsa) filter for speech synthesis. *Electronics and Communications in Japan (Part I: Communications)*, 66(2):10–18.
- Iriondo, I., Socoró, J. C., and Alías, F. (2007). Prosody modelling of spanish for expressive speech synthesis. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–821. IEEE.

- Iwahashi, N. and Sagisaka, Y. (1994). Speech spectrum transformation by speaker interpolation. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 1, pages I–461. IEEE.
- Jang, G.-J. and Cho, H.-Y. (2011). Efficient spectrum estimation of noise using line spectral pairs for robust speech recognition. *Electronics letters*, 47(25):1399–1401.
- Kain, A. and Macon, M. W. (1998). Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 285–288. IEEE.
- Karhila, R., Remes, U., and Kurimo, M. (2013). Hmm-based speech synthesis adaptation using noisy data: Analysis and evaluation methods. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6930–6934. IEEE.
- Käthner, I., Ruf, C. A., Pasqualotto, E., Braun, C., Birbaumer, N., and Halder, S. (2013). A portable auditory p300 brain–computer interface with directional cues. *Clinical neurophysiology*, 124(2):327–338.
- Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. *Proc. MAVEBA*, pages 13–15.
- Kim, S., Kim, J., and Hahn, M. (2006). Hmm-based korean speech synthesis system for handheld devices. *Consumer Electronics, IEEE Transactions on*, 52(4):1384–1390.
- King, S., Clark, R. A., Mayo, C., and Karaiskos, V. (2008). The blizzard challenge 2008. <http://www.festvox.org/blizzard/>.
- King, S. and Karaiskos, V. (2013). Blizzard challenge 2013. <http://www.festvox.org/blizzard/>.
- Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67:971.
- Klein, R., Kyrilov, A., and Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, pages 158–162. ACM.

- Kokkinakis, K. and Loizou, P. C. (2011). Evaluation of objective measures for quality assessment of reverberant speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2420–2423. IEEE.
- Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117.
- Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283.
- Krishnamurthy, J. and Mitchell, T. M. (2013). Vector space semantic parsing: A framework for compositional vector space models. *ACL 2013*, page 1.
- Krum, K. J., Conrad, M. J., Hulten, G. J., and Mendhro, U. A. (2014). User interface presenting an animated avatar performing a media reaction. US Patent 8,635,637.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.
- Latorre, J., Wan, V., Gales, M. J., Chen, L., Chin, K., Knill, K., and Akamine, M. (2012). Speech factorization for hmm-tts based on cluster adaptive training. In *INTERSPEECH*.
- Le Roux, J., Sagot, B., Seddah, D., et al. (2012). Statistical parsing of spanish and data driven lemmatization. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*.
- LeCun, Y. (2012). Learning invariant feature hierarchies. In *Computer vision—ECCV 2012. Workshops and demonstrations*, pages 496–505. Springer.
- Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech and language*, 9(2):171.
- Liang, H. and Dines, J. (2011). Phonological knowledge guided hmm state mapping for cross-lingual speaker adaptation. In *INTERSPEECH*, pages 1825–1828.

- Ling, Z.-H., Richmond, K., and Yamagishi, J. (2013). Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):207–219.
- Liu, F.-H., Stern, R. M., Huang, X., and Acero, A. (1993). Efficient cepstral normalization for robust speech recognition. In *Proceedings of the workshop on Human Language Technology*, pages 69–74. Association for Computational Linguistics.
- Llisterri, J. (1992). Speaking styles in speech research. In *Workshop on Integrating Speech and Natural Language*.
- Lopez Ludeña, V., San Segundo Hernández, R., Montero Martínez, J. M., Barra Chicote, R., and Lorenzo Trueba, J. (2012). Architecture for text normalization using statistical machine translation techniques. In *IberSPEECH 2012 - VII Jornadas en Tecnologia del Habla and III Iberian SLTech Workshop*. Telecomunicacion.
- Lorenzo-Trueba, J. (2012). Design and test of an expressive speech synthesis system based on speaker adaptation techniques. Master’s thesis, UPM-ETSIT.
- Lorenzo-Trueba, J., Barra-Chicote, R., Raitio, T., Obin, N., Alku, P., Yamagishi, J., and Montero, J. M. (2012a). Towards glottal source controllability in expressive speech synthesis. In *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13*.
- Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., and Montero, J. M. (2015). Emotion transplantation through adaptation in hmm-based speech synthesis. *Computer Speech & Language*.
- Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., and Montero, J. M. (2014). Towards cross-lingual emotion transplantation. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 199–208. Springer.
- Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., and Montero, J. (2013a). Evaluation of a transplantation algorithm for expressive speech synthesis. In *proceedings of Workshop en Tecnologias Accesibles, IV Congreso Español de Informatica CEDI2013*.
- Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., and Montero, J. M. (2013b). Towards speaking style transplantation in speech synthesis. In *8th ISCA Speech Synthesis Workshop*.

- Lorenzo-Trueba, J., Watts, O., Barra-Chicote, R., Yamagishi, J., King, S., and Montero, J. M. (2012b). Simple4all proposals for the albayzin evaluations in speech synthesis. In *Iber-speech2012, VII Jornadas en Tecnologia del Habla and III Iberian SLTech Workshop*.
- Lorenzo-Trueba, J., Watts, O., Barra-Chicote, R., Yamagishi, J., King, S., and Montero, J. M. (2012c). Simple4all proposals for the albayzin evaluations in speech synthesis. *Proc. Iber-speech*, 2012.
- Luo, Y. H.-L. and da Cruz, L. (2014). A review and update on the current status of retinal prostheses (bionic eye). *British medical bulletin*, page ldu002.
- Lutfi, S. B. L. (2013). *User-centric Need-driven Affect Modeling for Spoken Conversational Agents: Design and Evaluation*. PhD thesis, ETSIT-UPM.
- Lutfi, S. L., Fernández-Martínez, F., Lorenzo-Trueba, J., Barra-Chicote, R., and Montero, J. M. (2013). I feel you: The design and evaluation of a domotic affect-sensitive spoken conversational agent. *Sensors*, 13(8):10519–10538.
- Ma, J. and Loizou, P. C. (2011). Snr loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. *Speech Communication*, 53(3):340–354.
- Macon, M. W. and Clements, M. A. (1996). Speech concatenation and synthesis using an overlap-add sinusoidal model. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. On Conference Proceedings., 1996 IEEE International Conference - Volume 01, ICASSP '96*, pages 361–364, Washington, DC, USA. IEEE Computer Society.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.
- Mattar, E. (2013). A survey of bio-inspired robotics hands implementation: New directions in dexterous manipulation. *Robotics and Autonomous Systems*, 61(5):517–544.
- McHugh, A. (1976). Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. Technical report, DTIC Document.
- Mehrabian, A. (1980). *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*. Oelgeschlager, Gunn & Hain Cambridge, MA.

- Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., and Ranzato, M. (2014). Learning longer memory in recurrent neural networks. *arXiv preprint arXiv:1412.7753*.
- Moinet, A., Dutoit, T., and Latour, P. (2013). Audio time-scaling for slow motion sports videos. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, pages 314–320.
- Montero, J. and Barra-Chicote, R. (2012). The albayzin 2012 speech synthesis evaluation (albayzin 2012 ss). *Proc. Iberspeech*, 2012.
- Montero, J., Gutierrez-Arriola, J., Palazuelos, S., Enriquez, E., Aguilera, S., and Pardo, J. (1998). Emotional speech synthesis: From speech database to tts. In *Fifth International Conference on Spoken Language Processing*.
- Monzo, C., Iriondo, I., and Socoró, J. C. (2014). Voice quality modelling for expressive speech synthesis. *The Scientific World Journal*, 2014.
- Moon, T. K. (1996). The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60.
- Moreno-Sandoval, A., De la Madrid, G., Alcántara, M., Gonzalez, A., Guirao, J., and De la Torre, R. (2005). The spanish corpus. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam: John Benjamins Publishing Company, pages 135–161.
- Mousa, A. (2010). Voice conversion using pitch shifting algorithm by time stretching with psola and re-sampling. *Journal of electrical engineering*, 61(1):57–61.
- Murray, I. R. and Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4):369–390.
- Murray, I. R., Edgington, M. D., Campion, D., and Lynn, J. (2000). Rule-based emotion synthesis using concatenated speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.
- Nakano, Y., Tachibana, M., Yamagishi, J., and Kobayashi, T. (2006). Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis. In *INTER-SPEECH*.

- Narendra, N. and Sreenivasa Rao, K. (2013). Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis. *Applied Soft Computing*, 13(2):773–781.
- Narupiyakul, L., Sirinaovakul, B., and Cercone, N. (2013). Thai syllable analysis for rule-based text to speech system. *Computational Approaches to Assistive Technologies for People with Disabilities*, 253:139.
- Niwase, N., Yamagishi, J., and Kobayashi, T. (2005). Human walking motion synthesis with desired pace and stride length based on hsmm. *IEICE transactions on information and systems*, 88(11):2492–2499.
- Norris, A. E., Weger, H., Bullinger, C., and Bowers, A. (2014). Quantifying engagement: Measuring player involvement in human–avatar interactions. *Computers in Human Behavior*, 34:1–11.
- Nose, T., Kato, Y., and Kobayashi, T. (2007). Style estimation of speech based on multiple regression hidden semi-markov model. In *INTERSPEECH*, pages 2285–2288.
- Nose, T. and Kobayashi, T. (2012). An intuitive style control technique in hmm-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. *Speech Communication*.
- Obin, N. (2011). *MeLos: Analysis and Modelling of Speech Prosody and Speaking Style*. PhD thesis, Université Pierre et Marie Curie-Paris VI.
- Obin, N., Lanchantin, P., Lacheret, A., Rodet, X., et al. (2011). Discrete/continuous modelling of speaking style in hmm-based speech synthesis: Design and evaluation. In *Interspeech*.
- Obin, N., Roebel, A., Bachman, G., et al. (2014). On automatic voice casting for expressive speech: Speaker recognition vs. speech classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- Okubo, T., Mochizuki, R., and Kobayashi, T. (2006). Hybrid voice conversion of unit selection and generation using prosody dependent hmm. *IEICE - Trans. Inf. Syst.*, E89-D(11):2775–2782.
- Okuda, K., Kawahara, T., and Nakamura, S. (2002). Speaking rate compensation based on likelihood criterion in acoustic model training and decoding. In *INTERSPEECH*.

- Olive, J. (1980). A scheme for concatenating units for speech synthesis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80.*, volume 5, pages 568–571. IEEE.
- Olmos, R., León, J. A., Jorge-Botana, G., and Escudero, I. (2013). Using latent semantic analysis to grade brief summaries: A study exploring texts at different academic levels. *Literary and linguistic computing*, 28(3):388–403.
- Ortiz-Rosario, A. and Adeli, H. (2013). Brain-computer interface technologies: from signal to action. *Reviews in the neurosciences*, 24(5):537–552.
- O'Shaughnessy, D., Barbeau, L., Bernardi, D., and Archambault, D. (1988). Diphone speech synthesis. *Speech communication*, 7(1):55–65.
- Oura, K., Yamagishi, J., Wester, M., King, S., and Tokuda, K. (2012). Analysis of unsupervised cross-lingual speaker adaptation for hmm-based speech synthesis using kld-based transform mapping. *Speech Communication*, 54(6):703–714.
- Paice, C. D. (1994). An evaluation method for stemming algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–50. Springer-Verlag New York, Inc.
- Parrott, W. (2001). *Emotions in social psychology: Essential readings*. Psychology Press.
- Picart, B., Brognaux, S., and Drugman, T. (2013). Hmm-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation. In *Proceedings of the 8th Speech Synthesis Workshop (SSW8)*, pages 19–24.
- Picart, B., Drugman, T., and Dutoit, T. (2011). Continuous control of the degree of articulation in hmm-based speech synthesis. In *INTERSPEECH*, pages 1797–1800.
- Picart, B., Drugman, T., and Dutoit, T. (2014). Analysis and hmm-based synthesis of hypo and hyperarticulated speech. *Computer Speech & Language*, 28(2):687–707.
- Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., and Picheny, M. A. (2006). The ibm expressive text-to-speech synthesis system for american english. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1099–1108.

- Planet, S. and Iriondo, I. (2012). Comparative study on feature selection and fusion schemes for emotion recognition from speech. *International Journal of Interactive Multimedia & Artificial Intelligence*, 1(6).
- Planet, S. and Iriondo, I. (2013). Children emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features. *Cognitive Computation*, 5(4):526–532.
- Plumpe, M., Acero, A., Hon, H., and Huang, X. (1998). Hmm-based smoothing for concatenative speech synthesis. In *Fifth International Conference on Spoken Language Processing*.
- Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. Harper & Row New York.
- Polettini, N. (2004). The vector space model in information retrieval-term weighting problem. *Entropy*, pages 1–9.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.
- Qian, Y., Xu, J., and Soong, F. K. (2011). A frame mapping based hmm approach to cross-lingual voice transformation. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5120–5123. IEEE.
- Qin, L., Ling, Z.-H., Wu, Y.-J., Zhang, B.-F., and Wang, R.-H. (2006). Hmm-based emotional speech synthesis using average emotion model. In Huo, Q., Ma, B., Chng, E.-S., and Li, H., editors, *Chinese Spoken Language Processing*, volume 4274 of *Lecture Notes in Computer Science*, pages 233–240. Springer Berlin Heidelberg.
- Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (1988). *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ.
- Raitio, T., Suni, A., Vainio, M., and Alku, P. (2013). Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. *Computer Speech & Language*.
- Rodriguez-Losada, D., Matia, F., Galan, R., Hernando, M., Montero, J., and Lucas, J. (2008). Urbano, an interactive mobile tour-guide robot. *Advances in Service Robotics*. Ed. H. Seok. *In-Teh*, pages 229–252.
- Sainz, I., Erro, D., Navas, E., and Hernáez, I. (2011). A hybrid tts approach for prosody and acoustic modules. In *INTERSPEECH*, pages 333–336.

- San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., D'Haro, L., López-Ludeña, V., Sánchez, D., and García, A. (2012). Design, development and field evaluation of a spanish into sign language translation system. *Pattern Analysis and Applications*, 15(2):203–224.
- San Segundo, R., Montero, J. M., López-Ludena, V., and King, S. (2012). Detecting acronyms from capital letter sequences in spanish. In *INTERSPEECH*.
- Sankar, A. and Lee, C.-H. (1995). Robust speech recognition based on stochastic matching. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 121–124. IEEE.
- Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace, and Co.
- Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing*, pages 111–126. Springer.
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. (2013). Paralinguistics in speech and language-state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.
- Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: variances and strategies. *Affective Computing, IEEE Transactions on*, 1(2):119–131.
- Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., and Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1):15–25.
- Seltzer, M. L. and Acero, A. (2012). Factored adaptation using a combination of feature-space and model-space transforms. In *INTERSPEECH*.
- Serrano, J. I., del Castillo, M. D., Oliva, J., and Iglesias, A. (2011). The influence of stop-words and stemming on human text base comprehension. *Proceedings of the European Perspectives on Cognitive Science*.
- Shen, H.-P., Yeh, J.-F., and Wu, C.-H. (2011). Speaker clustering using decision tree-based phone cluster models with multi-space probability distributions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1289–1300.

- Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2002). Eigenvoices for hmm-based speech synthesis. In *INTERSPEECH*.
- Shinoda, K. and Lee, C.-H. (1997). Structural map speaker adaptation using hierarchical priors. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 381–388. IEEE.
- Shozakai, M. and Nagino, G. (2004). Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models. In *INTERSPEECH*.
- Singh, M. and Verma, K. (2013). Text to speech synthesis for numerals into punjabi language. Master’s thesis, COMPUTER SCIENCE AND ENGINEERING DEPARTMENT THAPAR UNIVERSITY PATIALA.
- Siohan, O., Myrvoll, T. A., and Lee, C.-H. (2002). Structural maximum a posteriori linear regression for fast hmm adaptation. *Computer Speech & Language*, 16(1):5–24.
- Spark Jones, K. (1972). A statistical interpretation of term importance in automatic indexing. *Journal of Documentation*, 28(1):11–21.
- Srikanth, R., Bajibabu, B., and Prahallad, K. (2012). Duration modelling in voice conversion using artificial neural networks. In *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, pages 556–559. IEEE.
- Steel, R. G. D., Torrie, J. H., et al. (1960). Principles and procedures of statistics. *Principles and procedures of statistics*.
- Stylianou, Y., Cappé, O., and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2):131–142.
- Suni, A., Raitio, T., Vainio, M., and Alku, P. (2012). The glotthmm entry for blizzard challenge 2012–hybrid approach. In *The Blizzard Challenge 2012 workshop*.
- Takaki, S., Sawada, K., Hashimoto, K., Oura, K., and Tokuda, K. (2013). Overview of nitech hmm-based speech synthesis system for blizzard challenge 2013. In *Blizzard Challenge 2013*.
- Takeda, S., Kabuta, Y., Inoue, T., and Hatoko, M. (2013). Proposal of a japanese-speech-synthesis method with dimensional representation of emotions based on prosody as well as voice-quality conversion. *International Journal of Affective Engineering*, 12(2):79–88.

- Theune, M., Meijs, K., Heylen, D., and Ordelman, R. (2006). Generating expressive speech for storytelling applications. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1137–1144.
- Tiomkin, S., Malah, D., Shechtman, S., and Kons, Z. (2011). A hybrid text-to-speech system that combines concatenative and statistical synthesis units. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1278–1288.
- Togneri, R. and Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *Circuits and Systems Magazine, IEEE*, 11(2):23–61.
- Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1315–1318. IEEE.
- Tokuda, K., Zen, H., and Black, A. (2002). An hmm-based speech synthesis system applied to english. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 227–230. IEEE.
- Toman, M., Pucher, M., and Schabus, D. (2013). Multi-variety adaptive acoustic modeling in hsmm-based speech synthesis. In *8th ISCA Speech Synthesis Workshop*.
- Toman, M. E. and Pucher, M. (2013). Structural kld for cross-variety speaker adaptation in hmm-based speech synthesis. In *Proc. SPPRA, Innsbruck, Austria*.
- Tonomura, M., Kosaka, T., and Matsunaga, S. (1995). Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 688–691. IEEE.
- Tonta, Y. and Darvish, H. R. (2010). Diffusion of latent semantic analysis as a research tool: A social network analysis approach. *Journal of Informetrics*, 4(2):166–174.
- Türk, O. and Schröder, M. (2010). Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. *Trans. Audio, Speech and Lang. Proc.*, 18(5):965–973.
- Valentini-Botinhao, C., Yamagishi, J., King, S., and Maia, R. (2014). Intelligibility enhancement of hmm-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion. *Computer Speech & Language*, 28(2):665–686.

- Veiga, A., Celorico, D., Proença, J., Candeias, S., and Perdigão, F. (2012). Prosodic and phonetic features for speaking styles classification and detection. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 89–98. Springer.
- Viswanathan, M. and Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer Speech & Language*, 19(1):55–83.
- Wander, J. D. and Rao, R. P. (2014). Brain–computer interfaces: a powerful tool for scientific inquiry. *Current Opinion in Neurobiology*, 25:70–75.
- Watts, O. (2012). *Unsupervised Learning for Text-to-Speech Synthesis*. PhD thesis, University of Edinburgh.
- Watts, O., Wu, Z., and King, S. (2015). Sentence-level control vectors for deep neural network speech synthesis. In *Proc. Interspeech*.
- Williamson, D. F., Parker, R. A., and Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921.
- Woodland, P. C., Leggetter, C., Odell, J., Valtchev, V., and Young, S. (1995). The 1994 htk large vocabulary speech recognition system. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 73–76. IEEE.
- Wu, C.-H., Hsia, C.-C., Liu, T.-H., and Wang, J.-F. (2006). Voice conversion using duration-embedded bi-hmms for expressive speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1109–1116.
- Wu, Y.-J., King, S., and Tokuda, K. (2008). Cross-lingual speaker adaptation for hmm-based speech synthesis. In *Chinese Spoken Language Processing, 2008. ISCSLP’08. 6th International Symposium on*, pages 1–4. IEEE.
- Wu, Y.-J., Nankaku, Y., and Tokuda, K. (2009). State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis. In *INTERSPEECH*, pages 528–531.
- Wu, Z., Valentini-Botinhao, C., Watts, O., and King, S. (2015). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4460–4464. IEEE.

- Wu, Z.-Z., Kinnunen, T., Chng, E., and Li, H. (2010). Text-independent f0 transformation with non-parallel data for voice conversion. In *INTERSPEECH*, pages 1732–1735.
- Wundt, W. (1897). *Outlines of psychology*. St. Clair Shores, MI: Scholarly Press.(CH Judd, transl., original 1897).
- Yaghoubzadeh, R., Kramer, M., Pitsch, K., and Kopp, S. (2013). Virtual agents as daily assistants for elderly or cognitively impaired people. In *Intelligent Virtual Agents*, pages 79–91. Springer.
- Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):66–83.
- Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). Hmm-based expressive speech synthesis-towards tts with arbitrary speaking styles and emotions. In *Proc. of Special Workshop in Maui (SWIM)*.
- Yamagishi, J., Masuko, T., and Kobayashi, T. (2007). A style control technique for hmm-based expressive speech synthesis. *IEICE transactions on information and systems*, 90(9):1406–1413.
- Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 88(3):502–509.
- Yamagishi, J., Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2003). A training method of average voice model for hmm-based speech synthesis. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 86(8):1956–1963.
- Yanagisawa, K., Latorre, J., Wan, V., Gales, M. J., and King, S. (2013). Noise robustness in hmm-tts speaker adaptation. *order*, 5:10.
- Yoshimura, T., Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2013). Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for hmm-based speech synthesis. *8th ISCA Speech Synthesis Workshop*.
- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1997). Speaker interpolation in hmm-based speech synthesis system. In *EUROSPEECH*.

- Zen, H., Braunschweiler, N., Buchholz, S., Knill, K., Krstulovic, S., and Latorre, J. (2010). Hmm-based polyglot speech synthesis by speaker and language adaptive training. In *Seventh ISCA Workshop on Speech Synthesis*.
- Zen, H., Keiichi, T., Takashi, M., Takao, K., and Tadashi, K. (2007). A hidden semi-markov model-based speech synthesis system. *IEICE TRANSACTIONS on Information and Systems*, 90(5):825–834.
- Zen, H., Nankaku, Y., and Tokuda, K. (2011). Continuous stochastic feature mapping based on trajectory hmms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(2):417–430.
- Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4470–4474. IEEE.
- Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7962–7966. IEEE.
- Zen, H., Tokuda, K., and Black, A. (2009). Review: Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.
- Zhang, H., Smith, L. C., Twidale, M., and Gao, F. H. (2011). Seeing the wood for the trees: Enhancing metadata subject elements with weights. *Information technology and libraries*, 30(2):75–80.
- Zovato, E., Pacchiotti, A., Quazza, S., and Sandri, S. (2004). Towards emotional speech synthesis: A rule based approach. In *Fifth ISCA Workshop on Speech Synthesis*.

Appendix A

Other Studies on Transplantation

A.1 Preliminary Speaking Styles Transplantation Evaluation

Our first transplantation evaluation that did not consider emotions attempted to transplant speaking styles learned from C-ORAL-Rom. In this evaluation what we wanted to measure was how different and simpler approaches to transplantation would fare for the task.

A.1.1 Alternatives to Transplantation: Copying the Speaking Style Average Model

In order to test the relevance of the proposed transplantation adaptation technique we defined a set of alternatives that could be considered for expressive synthesis, namely copying the different feature streams from the average models into the target speaker model. In this case it meant copying either the prosody features (F0 and duration streams) or the spectral features. This would not be an easy thing to do in a situation in which every voice was trained independently following the traditional HMM-based modeling, as the decision trees would not be shared. But because our voices share a common background model and the adaptation process keeps the trees intact, copying the prosody or the spectrum from the style averages is as simple as replacing the desired model files in the target speaker.

Figure A.1: Results of the perceptual test comparing the user preferences between transplantation and copy of prosody for all test speakers and styles on C-ORAL-Rom database.

Copying the Prosody

Prosody is known to carry a very significant portion of the expressive load of speech [Barra-Chicote, 2011]. As such it is assumable that by simply copying the prosody (only F0 and phone durations are considered) of a clear representative of a speaking style model would yield acceptable extrapolation results at least if the target speakers' F0 and the average's F0 are not too dissimilar. On the other hand, it is also possible that if this last condition does not hold the quality of the output voice would degrade or that instabilities might appear.

Copying the Spectrum

Spectrum is assumed to hold most of features related to the identifiability of the speaker, although it also includes some expressive features [Barra-Chicote, 2011]. In that sense it is safe to assume that this approach would not be very successful for extrapolating the expressiveness of the model but instead copy the identity of the style's average. This by itself does not seem useful for the task at hand, although for example in some extreme cases where a speaker can be clearly associated to a particular style, it would fulfill a similar purpose. Nonetheless, it seems interesting to test the results this kind of approach would give.

A.1.2 Perceptual Test Description

To test the effectiveness of the proposed technique we prepared a web-based perceptual test in which 32 listeners were asked to establish two different rankings in order of preference: adequacy

of the speaking style to the synthesized text and similarity to the original speaker. The decision to make the test ranking-based was taken because, as the task is considerably difficult and there is no natural voice reference available for the listener, comparing the system between themselves instead of assigning a value to them facilitated the testing process.

The target’s speaker neutral voice and the speaking style average model were synthesized to be added as top-line systems: the neutral voice would provide the top-line for the similarity analysis and the average model for the adequacy of the speaking style task.

The test consists of 4 systems (top-line system, transplantation system, copy-prosody system and copy-spectrum system) and 4 styles (news, sports, interviews and parliamentary speech) for each of the 3 evaluated target speakers. A description of the training data can be seen in table A.1.

Speaker	Amount of data
News	24 min
Politic	10 min
Sports	26 min
Interviews	26 min
UVD	2h
JEC	7 min
JLC	3 min

Table A.1: Description of the training data for the speaking styles transplantation evaluation. All speakers are male and speaking styles are extracted from the Spanish section of C-ORAL-Rom.

Regarding the interface, the listeners were presented with all five stimuli at the same time so that they could be played as many times as necessary. The stimuli distribution was designed following a balanced Latin square randomization of the questions pattern, resulting in a total of 8 utterances required per style. The utterances were extracted from real media texts and were selected so as to be longer than 10 words for the phrase-level prosody to become relevant. Finally, the minimum acceptable number of tests per target speaker was decided to be 16: two complete rounds of tests.

A.1.3 Results

An initial consideration verified by table A.2 is that copying the spectrum of the average or using the average itself is not a valid transplantation technique because these methods do not keep the identity of the target speaker, even if the perceived adequacy is comparable between the

Preference	Expressiveness	Similarity
Interview	38,9	41,1
News	48,9	44,4
Speech	58,9	65,6
Sports	46,7	58,9

Table A.2: Preference results in for the evaluated utterances in percentages in terms of expressiveness and similarity. The results show the preference for the transplanted system over the copy of prosody on C-ORAL-Rom database.

ADEQUACY	Read Speech	Style Average	Spectrum	Prosody	Transplantation
Sports	2.57	2.50	2.15	3.96	3.81
Interviews	2.63	3.69	2.98	2.94	2.76
News	2.98	3.22	2.68	3.37	2.76
Politic	3.49	2.82	2.40	3.02	3.26
Average	2.92	3.06	2.55	3.32	3.15

Table A.3: Results in adequacy in MOS scale for the different systems averaged between the 3 target speakers in the first speaking styles transplantation evaluation on C-ORAL-Rom database.

SIMILARITY	Read Speech	Average	Spectrum	Prosody	Transplantation
Sports	4.31	1.85	2.36	3.13	3.35
Interviews	4.50	1.68	2.28	3.37	3.18
News	4.36	1.89	2.12	3.41	3.22
Politic	4.40	1.73	2.35	3.01	3.51
Average	4.39	1.79	2.28	3.23	3.32

Table A.4: Results in similarity in MOS scale for the different systems averaged between the 3 target speakers in the first speaking styles transplantation evaluation on C-ORAL-Rom database.

5 proposed systems (table A.1.3). As such the analysis will focus only on when is adaptation considered by the listeners better than copy of prosody and vice versa.

From both table A.2 and figure A.1 we can see that there is no significant difference between transplantation and copying the prosody of the average model. In a global level it can be seen that political speech (but also live sports commentary) favor the transplantation-based system, while news favors the copy of prosody in the similarity department, without relevant differences in adequacy. This is reinforced by the similarity preference results of JEC, the high pitched speaker, for which the results definitely show that transplantation is the preferred technique.

The consideration to be drawn from the adequacy results is that even if in average the test results do not favor any of the techniques, the more different speaker-style pair of models (i.e. UVD

with politics or JLC with politics) also appear to be more adequate for transplantation than for copy of prosody. This could be seen as a hint that while copying the prosody is an acceptable method of extrapolating the speaking style, it stops being reliable when the pair of models is too disparate in prosody. On the other hand, transplantation through adaptation does not fall off in these kind of situations because not only the prosody is adapted but also the spectrum of the models, preventing instabilities or unnatural sounds from appearing.

Finally, it also became apparent that attempting to transplant expressiveness with such limited amounts of data, and particularly multi-speaker data was very difficult for our transplantation system. As such, we decided that we had to record our own single-speaker speaking styles database in order to work with different expressiveness.

A.2 SSS Transplantation Evaluation

A second speaking styles evaluation was carried out after the recording of the Spanish Speaking Styles (SSS) database (described in detail in section 5.2). In this case we compared out the proposed expressiveness transplantation system with traditional read speech synthesis in order to evaluate which system the listeners preferred.

A.2.1 Evaluation Design

The four speaking styles of SSS were evaluated (news, interviews, sports broadcasts and political speech) for both the proposed expressiveness transplantation system and the traditional STRAIGHT-based read speech system. 4 variables were evaluated, adequacy and similarity as a forced preference test and speech quality and expressive strength in a 5 points MOS scale. The preference question asked the evaluators to choose which system they preferred for conveying the target speaking style and the similarity question asked the evaluators to choose which of the two systems was more similar to the target speaker (i.e. they had to choose which resembled more to the target neutral speaker between the transplanted target speaker and the source neutral speaker).

The AVM was trained using the train set of SSS (described in table 5.1) and the three target male speakers considered for the emotion transplantation evaluation (see table 4.5). The model and vocoder configuration is the same as in the previous evaluations.

Preference	Adequacy	Similarity
Sports	94	81
Politic	89	75
Interviews	60(X)	78
News	55(X)	72
Average	75	77

Table A.5: Results in preference in percentage towards the proposed system for the adequacy and similarity in the SSS transplantation evaluation. The X represents a not statistically significant result.

Speech Quality	Transplanted	Neutral
Sports	3.31(X)	3.50(X)
Politic	3.23	3.86
Interviews	3.48(X)	3.52(X)
News	2.94	3.48
Average	3.24	3.54

Table A.6: Results in MOS scale for speech quality for both the proposed and neutral system in the SSS transplantation evaluation. A X represents a not statistically significant result.

A.2.2 Evaluation Results

In tables A.5 to A.7 we can see the results for the SSS transplantation evaluation. Regarding the adequacy and similarity preference results (table A.5) we can see how for similarity all the speaking styles significantly favor the proposed system, for an average of 77% preference, proving how the proposed expressiveness transplantation system is capable of maintaining the target speaker identity also when transplanting speaking styles. Regarding the preference results, both sports broadcasts and political speech significantly favor the transplanted system (94% and 89% preference respectively), while news and interviews show no preference between read speech and transplanted speech. This can be explained by how read speech and interviews can also be adequately conveyed with traditional read speech but both sports commentary and political speech can be vastly improved with a more expressive way of synthesizing. These results agree with those we will obtain in the complete system evaluation in section 5.3.4.

In terms of speech quality (table A.6), we can see how there is a significant but small reduction in speech quality for the proposed transplanted system of 0.30, even smaller than the reduction obtained in the emotion transplantation results.

Regarding the results for the perceived expressive strength (table A.7), we can see how the ex-

Expressive Strength	Transplanted	Neutral
Sports	3.75	2.56
Politic	3.38	2.80
Interviews	3.03(X)	2.95(X)
News	3.35	2.98
Average	3.38	2.82

Table A.7: Results in MOS scale for perceived expressive strength for both the proposed and neutral system in the SSS transplantation evaluation. A X represents a not statistically significant result.

pressiveness transplantation system is capable of providing significant improvements over read speech in all speaking styles excepting interviews, for an average of 0.56 increase. Particularly in live sports broadcasts we can provide a 1.2 increase in perceived expressive strength, although for political speech and news broadcasts we provide a 0.6 and 0.4 increase respectively.

All in all we can say how the proposed paralinguistic information transplantation system is also capable of correctly adding speaking styles information into a desired target speaker for whom expressive information is not available as long as we have the adequate source models for learning the paralinguistic information.

A.3 Cross-lingual Emotion Transplantation

The proposed cross-lingual emotion transplantation method is based on the state mapping principle [Oura et al., 2012]. The method begins by obtaining the mapping of the closest states amongst all states by means of the Karhunen-Loeve Divergence (KLD) [Toman and Pucher, 2013]. Then, the emotional adaptation function obtained by means of the transplantation method is converted from the source model and source language into the target model and target language. Finally we apply the emotional transformation to the target language. A flowchart overviewing the complete process can be seen in figure A.2.

A.3.1 Experimental Evaluation

Informal evaluations of the initial implementation of the cross-lingual emotion transplantation system showed that it is already capable of successfully transplanting the prosodic features between languages thanks to the state mapping technique, but a feature stream by feature stream

Figure A.2: Step by step block diagram of the emotion transplantation method. The spheres represent the speaker models and the thin arrows the adaptation transforms. The state mapping process is represented by the rectangle and the result of the mapping as the thick arrow. The two colors also represent the different languages.

transplantation showed that there is a significant amount of spectral noise introduced when transplanting the Cepstral information, especially in fundamentally prosodic emotions such as anger, which we will be using for the rest of the experimental evaluation section.

Figure A.3: Illustrations of the three different considered cross-lingual experimental variations. Black filled surfaces represent nonzero values.

In an attempt to quantify and lessen the effects of the introduced spectral noise we tried a number of different experiments: increase the minimum number of frames per adaptation node, reduce the adaptation covariance transformation matrix block size and reduce the covariance adaptation bandwidth. We also developed an objective method to compare the results of the different approaches based on a combination of KLD and Principal Component Analysis (PCA). KLD is

used to obtain the distances between all the speaker models used in the cross-lingual emotion transplantation process. PCA is used to project the KLD values into a bi-dimensional space that allows us to easily measure the distances in an objective way. This process is only done to the spectral features stream of the models as we want to measure the spectral distortion introduced to the process. In the ideal environment, the distance between the transplanted model (Korin-anger-S in the figures) and the target language model (Korin-neutral) should be the same as the distance between the source model (joa-anger-S) and the source language model (joa-neutral), which means obtaining a ratio of 1 in equation A.1. The Spanish source data is a subsection of the SEV database [Barra-Chicote et al., 2008], while the English target data was recorded in CSTR, University of Edinburgh. A description of the training data can be seen in table A.8

Speaker	Language	Amount of data
Korin-anger	English	33 min
Korin-neutral	English	1 h
joa-anger	Spanish	29 min
joa-neutral	Spanish	30 min

Table A.8: Description of the training data for the cross-lingual emotion transplantation. All speakers are male.

Adaptation Function Occupancy Threshold

The first approach increased the adaptation function occupancy threshold. The purpose of this is to reduce the size of the decision tree used for the adaptation process in the source language by increasing the number of frames that must be present for each node (represented graphically in the first graph of figure A.3). This produces more global transformation functions that convert the target language in a less complex fashion, potentially reducing the spectral noise introduced in the transplantation process. On the other hand, making use of more global adaptation functions means that there is less specificity in the transplantation, which could have an effect in the perceived emotion after transplanting.

Figure A.4: PCA projection of the different adaptation threshold experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.

The results can be seen in figure A.4, where it can be seen that while for lower threshold values there is no variation, for a very high value the distances between models clearly shrink, particularly between the emotional and neutral models of both languages. This means that using a higher threshold reduces the differences between the emotional and the neutral models.

Adaptation Variance Block Size

The second analysis aims to reduce the nonzero values in the adaptation function variance matrix by reducing the amount of intra-feature coefficients that influence each other (second illustration in figure A.3). That is, if in the standard block size every coefficient of a feature stream is able to influence each other, a smaller block size prevents the higher order coefficients from affecting lower ones.

The projection of the models obtained by manipulating the size of the blocks in the adaptation matrix can be seen in figure A.5, and they show that both a big block size and a very small block size do not provide the optimal results, while an intermediate value provides less distant models. This is because a too small block size removes too many coefficients in the adaptation variance matrix, which results in a very coarse adaptation, unable to deal correctly with emotional data [Yamagishi et al., 2009].

Figure A.5: PCA projection of the different adaptation block size experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.

Adaptation Variance Bandwidth

The final variant controls the values in the adaptation function variance matrix by only allowing a certain bandwidth of them around the diagonal to be nonzero as seen in the third illustration in figure A.3. This approach also aims to reduce the interaction between higher and lower order coefficients in the adaptation of the feature streams, and it is supposedly smoother because there are no abrupt cuts in which coefficients affect each other. This approach makes special sense when using Linear Spectral Pairs (LSP) as the spectral features instead of Cepstral features, as the order of the coefficient correlates with the central frequency of the LSP. Thus, this approach directly limits the bandwidth that is considered in the adaptation process.

In the case of controlling the bandwidth of the adaptation variance matrix (figure A.6), we can see that the lower the adaptation bandwidth the closer all the models become to each other. This translates into less noisy transplanted models but at the same time less expressive and identifiable models. This is also expected to give less expressive models as in the block size case.

Figure A.6: PCA projection of the different adaptation bandwidth experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system, constant across experiments.

A.3.2 Perceptual Evaluation and Covariance Analysis

In an attempt to verify the validity of the objective measures, we carried out a small perceptual evaluation which aimed to obtain subjective speech quality results for all the considered systems. In this evaluation we presented the listeners with 14 different audio samples (one per system and only one at a time), and asked them to rate the perceived speech quality from 1 (very bad) to 5 (very high). A total of 14 utterances were synthesized according to the Latin square evaluation strategy in order to remove any bias in the evaluation process. Finally, we measured the correlation between a ratio of the model distances and the speech quality as follows:

$$\text{corr}\left(\frac{\text{dist}(\text{TransplantedEnglish}_i - \text{NeutralEnglish})}{\text{dist}(\text{EmotionalSpanish}_i - \text{NeutralSpanish})}, \text{SpeechQuality}_i\right) \quad (\text{A.1})$$

In the results of measuring said correlation (table A.9) we can see that there is a very strong relationship between the ratio of the distances between transplanted and source models and the perceived speech quality. This means that the proposed objective measure is a good tool for measuring how much we can improve speech quality compared to the initial cross-lingual transplantation system.

Evaluation System	Threshold	Block Size	Bandwidth
Measured Correlation	-0.948	-0.952	-0.893

Table A.9: Covariance analysis results for the three cross-lingual experimental environments.

A.4 Transplantation System Performance Analysis

The biggest limitation of the proposed paralinguistic features transplantation method is that it slightly decreases the synthesized speech quality by introducing some artifacts. In an attempt to predict why and in which contexts those artifacts appeared, we developed an upgraded version of the state mapping algorithm that measures the distances between exactly the same contexts of two models. This means that in order to apply it we have to compare two parallel models trained with the same data, which is exactly what we have in our emotional data.

The main motivation behind this approach is that, based on the results we obtained from the cross-lingual transplantation evaluation we know that there is a correlation between distance between models and speech quality. As such, we expected that by measuring the distances between the contexts of the emotional model of the source speaker and the same contexts of the neutral model of the same speaker we would be able to predict when the adaptation function that will be transplanted will have errors, and thus induce artifacts.

A.4.1 Evaluating the distance between parallel contexts

First of all we had to modify the HTS code so that the state mapping would look for parallel contexts instead of looking for the states with the lowest KLD. Once that was done, we had to obtain the distances at state level for every context of the model separated by streams. Then, in order to better analyze the problematic contexts we split them into three groups: high distance, medium distance and low distance. The separation was done according to the following two thresholds:

$$HT = (iqr * 5) + upperq \quad (A.2)$$

$$LT = (iqr * 1) + upperq \quad (A.3)$$

Where states with distance higher than HT are classified as high distance states and those with a distance between HT and LT are classified as medium distance states. Iqr is the interquartile range of the distribution and $upperq$ the threshold value for the third quartile. This thresholding was applied for each of the 5 states of the 3 feature streams and its derivatives for every context.

The evaluation was carried out by analyzing a single pair of models: the neutral and happy models of joa. In order to detect potential errors we listened to each context once and labeled them as problematic or not. Due to the sheer amount of contexts and the laborious nature of the manual perceptual analysis only the contexts for the logF0 and mcep streams and for those, only the high distance and medium distance contexts (i.e. 1791 contexts in 337 short utterances for both neutral and happy models). The utterances were extracted from the training set and synthesized using the neutral and happy models.

Distance Group	Neutral	Happy
Mcep-high	15±7	18±8
Mcep-mid	9±6	7±5
LF0-high	26±5	23±4
LF0-mid	12±6	5±4

Table A.10: Percentage of contexts, with the error margin for a 95% confidence, that presented errors by feature stream and context distance.

The results, which can be seen in table A.10, show that while this approach is not a good predictor by itself, it is capable of showing that there is a significant difference in the amount of erroneous contexts when comparing the high distance group and medium distance group of LF0 and a difference albeit not significant in the case of the mcep groups. This means that even if context-to-context distances are not capable by themselves to predict artifacts in the contexts, they could prove helpful to that purpose if more effort is paid to the problem.

A.5 MRHMM-based transplantation

As part of the development of the discrete system introduced in section 5.4, we developed a system that included both an emotion control factor and a speaker control factor so that, as long as we have multiple speakers in the training model and multiple emotions for one of them, we can produce transplanted emotional speech with a target speaker for whom prior emotional information is not available.

Figure A.7: Flowchart of the proposed MRHMM-based emotion transplantation system.

The training process of the proposed method can be seen in figure A.7 and can be explained in the following steps:

1. Train the Background Emotional Model (BEM) through Speaker Adaptive Training (SAT) [Anastasakos et al., 1997] with both the source and target speakers' data and the source emotion data.
2. Codify the speaker and emotion multi-factor control vectors $cv = \{spk_i, emo_j\}$.
3. Re-estimate the BEM by means of the control vector and MR-HSMM to obtain the speaker and emotion independent model.

Once the speaker and emotion independent model is trained, the synthesis just requires the text labels and the target factors' vector with the desired output emotion and speaker to generate the acoustic parameters that can be used for the speaker generation.

Our current implementation considers a very simple BEM, where only the source speaker in the neutral emotion and the target emotion and the target speaker in the neutral emotion are considered, as we can see in table A.11. This allows us to control more closely any potential problems that can arise due to the multi- factor multiple regression, although we expect the system to be capable of handling additional layers of factorization such as speaker gender or multiple expressiveness such as Lombard speech on top of the current target speaker and emotion factors.

BEM	Data
Anger	ROS-neutral, ROS-anger, UPC-neutral
Happiness	ROS-neutral, ROS-happiness, UPC-neutral
Neutral	ROS-neutral, UPC-neutral
Sadness	ROS-neutral, ROS-sadness, UPC-neutral
Surprise	ROS-neutral, ROS-surprise, UPC-neutral

Table A.11: Speech data used to train each background emotion model in the MR-HMM based transplantation.

A.5.1 Perceptual Evaluation

Both emotion identification and similarity were evaluated in order to quantify the capabilities of the proposed system to adequately transplant the desired emotion while maintaining the target speaker identity. Emotion identification was evaluated by means of a multiple choice test, where the listeners were asked to select the emotion they perceived in the samples under evaluation (6 choices: anger, happiness, neutral, sadness, surprise and other). Similarity was measured by means of a forced preference test, where the evaluator was presented with a pair of reference samples (neutral samples from the source and target speaker) and was asked to select which of the two samples was more similar to the sample under evaluation.

Regarding the evaluation design, four emotions (anger, happiness, sadness and surprise) learned from ROS and two neutral speakers (UPC from INTERFACE database [Hozjan et al., 2002] and ROS) were considered for a total of 10 systems to be analyzed, about 30 minutes of speech were used for each speaker-emotion pair. The speaker models were trained using the same configuration as all previous evaluations (see section 4.3). Each emotion was learned in the same triplet: UPC-neutral, ROS-neutral and ROS-emotion as can be seen in table A.11, and the emotion control vector factor was fixed at 1.00 for ROS and at 0.75 in the first evaluation and at 1.00 in the second for UPC. The evaluation was designed according to the Latin-square strategy [Gao, 2005] in order to obtain fully randomized and unbiased evaluation sets. This meant using 10 evaluation texts, which were selected from a test set of the SEV database and are known to be neutral in context so as not to bias the emotion identification task. A total of 100 samples were synthesized.

The test itself was carried out by means of a web interface. For each of the 10 samples included in the test, the evaluators were presented first with the target sample, that could be played at any time, and were asked to select the identified emotion. At the same time they were asked to evaluate similarity to two reference neutral samples, that were the neutral synthetic versions for

Src. Speaker	Factor	N	A	H	S	U	O
Neutral	0.0	69					18
Anger	1.0		88				
Happiness	1.0			41		20	
Sadness	1.0				75		
Surprise	1.0			24		54	

Table A.12: Confusion matrix in percentage of the emotion identification rates for the source speaker (ROS) in percentage for the MR-HMM based transplantation. A stands for anger, H for happiness, N for neutral, S for sadness, U for surprise and O for other. Results below the random threshold are omitted.

both source and target speaker of the sample under evaluation. Both source and target speakers were analyzed in order to measure the differences introduced by the transplantation process. Synthetic speech was used in the similarity section because we wanted to measure if there is any speaker identity degradation introduced by the emotion transplantation process.

A.5.2 Results

Emotion Identification Rates

First of all, we evaluated the emotion identification rates (EIR) for both emotion factors in the control vector and for both source and target speakers. The results can be seen in tables A.12 and A.13 respectively.

The emotion recognition task is known to be inherently difficult. Even so, the EIR obtained with the MRHMM-based modeling synthetic voice of the source speaker (ROS) are comparable to those obtained with standard HMM-based modeling [Barra-Chicote et al., 2010]. Most notably, happiness has a significant confusion with surprise, which causes them to have a low accuracy. The happiness results also show that the regression with the proposed factors coding needs improvements due to the randomness in its identification. A linear regression modeling has not been enough to model the complex nature of happiness [Barra-Chicote et al., 2010].

For the source speaker results (table A.12) we can see how the multiple regression modeling has been capable of capturing successfully anger and sadness, the negative emotions, for 88% and 75% recognition rates respectively. On the other hand there is a significant confusion between the happy emotions resulting in 41% recognition rates for happiness and 54% for surprise, which

Tgt. Speaker	Factor	N	A	H	S	U	O
Neutral	0.0	66					
	0.0	86					
Anger	0.75	31			21		
	1.0		43		33		
Happiness	0.75	31		21			21
	1.0			38	19		
Sadness	0.75				86		
	1.0				90		
Surprise	0.75	24				48	
	1.0					62	

Table A.13: Confusion matrix in percentage of the emotion identification rates for the target speaker (UPC) in percentage for the MR-HMM based transplantation. Column headers are analogous to table A.12.

are well above the random threshold in any case. The same effect was observed in previous evaluations for ROS voice (i.e. interpolation results in section 3.3.2). Neutral voice presents reasonable results at 70% recognition rates, with some confusion with the "other" category.

Looking at the transplanted speaker (UPC) EIR in table A.13, for the 0.75 emotion control factor we can see an very significant relative degradation in anger results compared to the source speaker (ROS). The 0.75 control factor value was not enough to express anger, as it was confused with the neutral emotion (31%). Increasing the control factor value to 1.0 makes anger partially recognizable. Also, the negative polarity of anger makes it consistently confusable with the other negative emotion, sadness (21% and 33% for both factor values respectively). These results hint that we should use an emotion control factor higher than 1.0 to increase anger identifiability. Prosodic emotions (sadness and surprise) have been successfully transplanted (118% and 100% relative EIR respectively). In this case, an emotion control factor of 0.75 is enough to attain high accuracy rates in contrast with the segmental and mixed-nature emotions (anger and happiness) that needed at least a control factor of 1.0.

Similarity Preference

Secondly, we evaluated the similarity of the transplanted speech with both the target speaker and the source speaker by means of a forced preference test. The results for the first evaluation for each emotion can be seen in figure A.8. These results show that, for the emotion control factor of 0.75, there is in general not a problem for recognizing the target speaker, although the speaker

Figure A.8: Similarity results in percentage of the first perceptual MR-HMM based transplantation evaluation with the transplanted emotion control factor set at 0.75.

similarity for the target speaker in the positive emotions shows that they are indistinguishable. This is likely because of the higher pitch that is associated to positive emotions, making the source speaker sound more similar to the target speaker.

The results for the second evaluation for each emotion can be seen in figure A.9. In this case we can see how in general the speaker similarity results do not change significantly due to the increase in the emotion control factor. In most cases the similarity preference towards the target speaker does not become smaller, which shows that speaker identity does not become distorted even for a 1.0 emotion control factor.

In general it is clear from the similarity results that the system is also capable of maintaining the identity of the target speaker for reasonable emotion control factors, while at the same time being capable of transplanting the desired emotion. In any case, the proposed transplantation system based in adaptation provides much better results in all the evaluated aspects.

Figure A.9: Similarity results in percentage of the second perceptual MR-HMM based transplantation evaluation with the transplanted emotion control factor set at 1.0.

Appendix B

Software and Versions

All the development has been carried out on Debian GNU/Linux and the toolkits with the versions specified in the next sections.

B.1 Speech Synthesis

- HTS 2.1.1 for HTK 3.4.1 - <http://hts.sp.nitech.ac.jp/>
- VCTK and VCTK2 - <http://homepages.inf.ed.ac.uk/jyamagis/software/page37>
- STRAIGHT Vocoder - http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html
- MR-HMM implementation for HTS 2.1.1 was kindly provided by Simon King and Oliver Watts of CSTR, UEDIN - <http://www.cstr.ed.ac.uk/>

B.2 Genre Prediction

- Python 2.7.3 - <https://www.python.org/download/releases/2.7.3/>
- numpy 1.6.1 - <https://pypi.python.org/pypi/numpy/1.6.1>
- scipy 0.9.0 - <https://pypi.python.org/pypi/scipy/0.9.0>