Tesis Doctoral

Síntesis de voz aplicada a la traducción voz a voz

Pablo Daniel Agüero

Director de Tesis:
Antonio Bonafonte Cávez

TALP Research Centre, Speech Processing Group
Departamento de Teoría de la Señal y Comunicaciones
Universidad Politécnica de Cataluña (UPC)

Barcelona, 2012

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Oficina de Doctorat
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Acta de qualificació de tesi doctoral

Curs acadèmic:

Nom i cognoms

DNI / NIE / Passaport

Programa de doctorat

Unitat estructural responsable del programa

Resolució del Tribunal

Reunit el Tribunal designat a l'efecte, el doctorand / la doctoranda exposa el tema de la seva tesi doctoral titulada

_____

_____.

Acabada la lectura i després de donar resposta a les qüestions formulades pels membres titulars del tribunal, aquest atorga la qualificació:

APTA/E NO APTA/E

(Nom, cognoms i signatura) President/a	(Nom, cognoms i signatura) Secretari/ària	
(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal

_____, _____ d'/de _____ de _____

El resultat de l'escrutini dels vots emesos pels membres titulars del tribunal, efectuat per l'Escola de Doctorat, a instància de la Comissió de Doctorat de la UPC, atorga la MENCIÓ CUM LAUDE:

SI NO

(Nom, cognoms i signatura) Presidenta de la Comissió de Doctorat	(Nom, cognoms i signatura) Secretària de la Comissió de Doctorat

Barcelona, _____ d'/de _____ de _____

A mi viejo

Resumen

Dentro de las tecnologías del habla, la conversión texto a voz consiste en la generación, por medios automáticos, de una voz artificial que genera idéntico sonido al producido por una persona al leer un texto en voz alta. En resumen, los conversores texto a voz son sistemas que permiten la conversión de textos en voz sintética.

El proceso de conversión texto a voz se divide en tres módulos básicos: procesamiento del texto, generación de la prosodia y generación de la voz sintética. En el primero de los módulos se realiza la normalización del texto (para expandir abreviaciones, convertir números y fechas en texto, etc), y en ocasiones, luego también se hace un etiquetado morfosintáctico. A continuación se procede a la conversión de los grafemas en fonemas y a la silabificación para obtener la secuencia de fonemas necesaria para reproducir el texto. Posteriormente, el módulo de prosodia genera la información prosódica para poder producir la voz. Para ello se predicen las frases entonativas y la entonación de la oración, y también la duración y la energía de los fonemas, etc. La correcta generación de esta información repercutirá directamente en la naturalidad y expresividad del sistema. En el último modulo de generación de la voz es donde se produce la voz considerando la información provista por los módulos de procesamiento del texto y prosodia.

El objetivo de la presente tesis es el desarrollo de nuevos algoritmos para el entrenamiento de modelos de generación de prosodia para la conversión texto a voz, y su aplicación en el marco de la traducción voz a voz. En el caso de los algoritmos de modelado de entonación, en la literatura se proponen generalmente enfoques que incluyen una estilización previa a la parametrización. En esta tesis se estudiaron alternativas para evitar esa estilización, combinando la parametrización y la generación del modelo de entonación en un todo integrado. Dicho enfoque ha resultado exitoso tanto en la evaluación objetiva (usando medidas como el error cuadrático medio o el coeficiente de correlación Pearson) como en la subjetiva. Los evaluadores han considerado que el enfoque propuesto tiene una calidad y una naturalidad superiores a otros algoritmos existentes en la literatura incluidos en las evaluaciones, alcanzando un MOS de naturalidad de 3,55 (4,63 para la voz original) y un MOS de calidad de 3,78 (4,78 para la voz original).

En lo referente al modelado de la duración se estudió la influencia de los factores segmentales y suprasegmentales en la duración de los fonemas. Con los resultados de este estudio se propusieron algoritmos que permiten combinar la información segmental y suprasegmental para realizar una predicción de la duración de los fonemas, tal como se propuso en otras publicaciones del tema en cuestión. A través de un estudio de los datos de entrenamiento se demostró la dependencia entre la duración de la sílaba y el número de segmentos constituyentes. Como consecuencia de estas observaciones, se propuso el modelado segmental utilizando la duración silábica, sin considerar una isocronía silábica

estricta.

Los primeros algoritmos propuestos consideran que la duración segmental puede modelarse como una fracción de la duración silábica. En consecuencia, cada segmento variará en función de la duración suprasegmental, ajustándose todos los fonemas constituyentes a la duración predicha de la sílaba. Sin embargo, la observación de la correlación entre la duración de la sílaba y la duración segmental nos permitió determinar que en algunas ocasiones pueden considerarse como fenómenos que no guardan una relación lineal entre ellos. Teniendo en cuenta esto, en esta tesis también se propuso el modelado de la duración segmental de manera condicional, considerándola como una fracción de la duración silábica, o bien en forma absoluta, independiente de la duración suprasegmental. Estos algoritmos propuestos utilizan una extrapolación para el modelado de la duración del enfoque planteado para el modelado de la entonación. La evaluación subjetiva sugiere que la predicción de la duración segmental en base a la duración de la sílaba usando duraciones relativas y absolutas alcanzan un MOS de naturalidad de 4,06 (4,59 para la voz original) y un MOS de calidad de 4,25 (4,65 para la voz original).

Finalmente, también se realizó un análisis de diversos modelos de junturas terminales usando tanto palabras como grupos acentuales: árboles de clasificación (CART), modelos de lenguaje (LM) y transductores de estados finitos (FST). La utilización del mismo conjunto de datos para los experimentos permitió obtener conclusiones relevantes sobre las diferencias de los diferentes modelos. Los experimentos realizados revelan la ventaja de la utilización de modelos de lenguaje a través de n-gramas (CART+LM) sobre el algoritmo más simple que predice junturas usando solamente CART. Tanto en el modelado usando palabras como grupos acentuales, CART+LM y FST resultaron superiores a la utilización de árboles de clasificación en forma aislada. Además, en todos los casos CART+LM resultó superior a FST debido a la posibilidad de utilizar información contextual más compleja a través de la probabilidad modelada con el árbol de clasificación, tales como etiquetas morfosintácticas adyacentes y la distancia a signos de puntuación.

Uno de los objetivos de esta tesis era mejorar la naturalidad y expresividad de la conversión texto a voz utilizando la prosodia del hablante fuente disponible en el proceso de traducción voz a voz como información adicional. Por ello se han desarrollado una serie de algoritmos para la generación de la prosodia que permiten la integración de la información adicional en la predicción de la entonación, la duración de los fonemas y la ubicación de junturas terminales.

Los diferentes modelos prosódicos de entonación, duración segmental y junturas terminales desarrollados en la primera parte de la tesis se adaptaron para incluir información prosódica extraída del hablante fuente. El objeto era mejorar la generación de la prosodia en la conversión texto a voz en el marco de la traducción voz a voz en aspectos tales como naturalidad, expresividad y consistencia con el estilo del hablante fuente.

En ese sentido esta tesis exploró diferentes enfoques para la transferencia de la entonación de un idioma a otro. Para ello se consideró la posibilidad de utilizar esquemas de anotación existentes, tales como ToBI o INTSINT. De esta manera, una vez obtenida la anotación de ambos idiomas, sería posible aplicar técnicas de aprendizaje automático para encontrar relaciones entre las anotaciones. Sin embargo, la conclusión fue que en este tipo de esquemas de anotación de eventos tonales se realizan ciertas suposiciones, tales como una discretización taxativa de los contornos, que pueden forzar el ajuste del fenómeno al esquema de anotación, y no viceversa, que es lo deseado. Esto puede llevar a una ano-

tación deficiente de los eventos tonales, y la utilización de esta información errónea solo conduciría a resultados pobres en la transferencia de la entonación.

Por ello se decidió la utilización de un enfoque de agrupamiento automático que permita encontrar un cierto número de tipos de movimientos tonales relacionados en los dos idiomas sin utilizar ninguna suposición acerca de su número. De esta manera, es posible utilizar esta codificación (obtenida luego del agrupamiento automático) de los contornos tonales del idioma origen como característica adicional en el modelado de la entonación del idioma destino. Los resultados experimentales demostraron la mejora introducida en el modelado de la entonación debido al enfoque propuesto, en comparación con un sistema base que no utiliza la información de la codificación del contorno del idioma origen. La mejora es importante en idiomas cercanos, tales como español y catalán. En el caso del español y el inglés, los resultados fueron apenas ligeramente mejores, debido en parte a la raiz diferente de los idiomas: latina y germánica respectivamente.

Si bien se decidió no realizar una transferencia de la duración segmental entre idiomas, en esta tesis se propuso transferir el ritmo del idioma origen al destino. Para ello se propuso un método que combina la transferencia del ritmo y la sincronización entre audios. Este último aspecto fue considerado debido al uso de la tecnología de traducción voz a voz en conjunción con video. Coordinar los aspectos gestuales con la voz traducida es importante a causa de los múltiples canales involucrados en la comunicación humana. En los experimentos se puedieron observar errores de sincronización muy bajos, cercanos a los 150 milisegundos, que convierte al enfoque propuesto en apto para su uso en sincronización de audio/video.

Por último, en esta tesis también se propuso una técnica de transferencia de pausas en el marco de la traducción voz a voz, mediante la utilización de información sobre alineamiento. El estudio de los datos de entrenamiento utilizando dos tipos diferentes de unidades de traducción, palabras y tuplas, arrojó como resultado la ventaja del uso de la última para dicha tarea. La tupla permite agrupar es su interior palabras que presentan un ordenamiento entre idiomas. En consecuencia, es posible transferir las pausas de un idioma a otro cuando estas se encuentran en la frontera de las tuplas. Una limitación importante de este enfoque es la imposibilidad para trasladar una pausa de una tupla de un idioma a otro, si esta se encuentra dentro de la misma. Para compensar esta deficiencia el algoritmo realiza una predicción de pausas adicionales utilizando algoritmos convencionales (CART, CART+LM, FST), teniendo en cuenta las pausas ya predichas mediante la transferencia de pausas entre idiomas.

Índice general

1. Introducción **1**
- 1.1. Arquitectura de un sistema de traducción voz a voz . . . 3
  - 1.1.1. Reconocimiento automático del habla . . . 4
  - 1.1.2. Traducción automática . . . 6
  - 1.1.3. Conversión texto a voz . . . 9
- 1.2. Proyectos relacionados con la traducción voz a voz . . . 12
- 1.3. TC-STAR . . . 16
  - 1.3.1. Resultados obtenidos en ASR . . . 16
  - 1.3.2. Resultados obtenidos en MT . . . 17
  - 1.3.3. Objetivos en TTS . . . 18
- 1.4. Objetivos de la tesis . . . 18
- 1.5. Estructura de la tesis . . . 19

2. Modelado prosódico en los sistemas de síntesis de voz **21**
- 2.1. Conversión de texto en habla . . . 22
  - 2.1.1. Procesamiento del texto . . . 22
  - 2.1.2. Modelado prosódico . . . 26
  - 2.1.3. Generación de voz artificial . . . 27
  - 2.1.4. Importancia de la prosodia en la generación de voz . . . 32
- 2.2. Entonación . . . 33
  - 2.2.1. Unidades de la entonación . . . 35
  - 2.2.2. La entonación en la conversión texto-voz . . . 36
  - 2.2.3. Modelos de entonación fonológicos . . . 37
  - 2.2.4. Modelos de entonación perceptuales . . . 38
  - 2.2.5. Modelos de entonación por estilización acústica superposicionales y no superposicionales . . . 38
- 2.3. Duración . . . 47
  - 2.3.1. Factores que influyen en la variación de la duración segmental . . . 47
  - 2.3.2. Generación de la duración en los TTS . . . 48
  - 2.3.3. Modelado de la duración usando suma de productos . . . 49
  - 2.3.4. Modelado de la duración usando CART . . . 50
  - 2.3.5. Modelado de la duración usando redes neuronales . . . 51
  - 2.3.6. Modelado segmental y suprasegmental . . . 51
- 2.4. Junturas terminales . . . 52
  - 2.4.1. Modelado de las junturas terminales . . . 53

2.4.2. Modelado de las junturas terminales usando CART 54
2.4.3. Modelado de las junturas terminales usando Bayes 54
2.4.4. Modelado de las junturas terminales usando redes neuronales 55
2.4.5. Otros algoritmos propuestos para el modelado de las junturas terminales 55
2.5. Conclusiones 56
2.5.1. Entonación 56
2.5.2. Duración segmental 56
2.5.3. Junturas terminales 57

3. Aportaciones en el modelado prosódico **59**
3.1. Modelado de la entonación 59
3.1.1. Problemas de la parametrización 60
3.1.2. El enfoque de parametrización y entrenamiento conjuntos (JEMA). . 63
3.1.3. Modelado de la entonación basado en curvas de Bézier 68
3.1.4. Modelado de la entonación usando el enfoque de Fujisaki 73
3.2. Modelado de la duración 76
3.2.1. Predicción de la duración usando dos niveles. 77
3.2.2. Modelado de la duración segmental como una fracción de la duración suprasegmental usando estimación separada 79
3.2.3. Modelado de la duración segmental como una fracción de la duración suprasegmental usando estimación conjunta 81
3.2.4. Modelado mixto de la duración segmental como una fracción de la duración suprasegmental y en forma absoluta usando estimación conjunta 82
3.3. Modelado de las junturas terminales 84
3.3.1. Modelado de las junturas terminales usando CART. 84
3.3.2. Modelado de las junturas terminales usando CART y un modelo de lenguaje. 85
3.3.3. Modelado de las junturas terminales usando transductores de estados finitos. 86
3.3.4. Modelado de las junturas terminales usando grupos acentuales. . . . 89
3.4. Conclusiones 89
3.4.1. Entonación 90
3.4.2. Duración 90
3.4.3. Junturas terminales 91

4. Validación experimental de las aportaciones **93**
4.1. JEMA: una prueba de concepto 93
4.1.1. Datos experimentales 94
4.1.2. Resultados experimentales 96
4.2. Validación de JEMA para el modelado de la entonación 97
4.2.1. Datos experimentales 98
4.2.2. Resultados experimentales 99
4.3. Validación de JEMA para el modelado de la duración 103
4.3.1. Datos experimentales. 103

4.3.2. Resultados experimentales. . . . 105
4.4. Experimentos sobre modelado de junturas terminales . . . 108
4.4.1. Datos experimentales. . . . 108
4.4.2. Resultados experimentales. . . . 109
4.5. Conclusiones . . . 111
4.5.1. Entonación . . . 111
4.5.2. Duración . . . 112
4.5.3. Junturas terminales . . . 112

5. Transferencia de la prosodia en la traducción oral 115
5.1. Limitaciones para la generación de la prosodia en un sistema de conversión texto a voz . . . 115
5.2. Generación de la prosodia en un sistema de traducción voz a voz . . . 119
5.3. Generación de la entonación utilizando la información de la fuente . . . 121
5.3.1. Corpus orales para la investigación en generación de prosodia en traducción . . . 123
5.3.2. Transferencia de información del contorno origen para generar el contorno destino . . . 125
5.3.3. Sistemas de anotación simbólica de la entonación . . . 126
5.3.4. Anotación de la entonación del hablante fuente . . . 129
5.3.5. Validación experimental . . . 134
5.4. Generación de la duración utilizando información de la fuente . . . 140
5.4.1. Influencia del ritmo en las unidades del habla . . . 141
5.4.2. Transferencia del ritmo entre idiomas. . . . 143
5.4.3. Sincronización de los audios de dos idiomas. . . . 145
5.5. Generación de pausas usando información de la fuente . . . 148
5.5.1. Transferencia de pausas usando tuplas. . . . 149
5.5.2. Condiciones experimentales . . . 151
5.5.3. Resultados experimentales . . . 151
5.6. Conclusiones . . . 153

6. Conclusiones y direcciones futuras 157

Apéndice A - Ogmios: el conversor texto a voz de la UPC 163
A.1. Procesamiento del texto . . . 164
A.2. Generación de la prosodia . . . 165
A.3. Generación de la voz . . . 166
A.4. Construcción de la voz sintética . . . 167

Apéndice B - Herramientas estadísticas utilizadas 169
B.1. Error cuadrático medio . . . 169
B.2. Coeficiente de correlación Pearson . . . 170
B.3. Box-plots . . . 172
B.4. Wilcoxon test . . . 173

Apéndice C - Corpus TC-STAR **175**
C.1. Corpus monolingüe 175
C.2. Corpus bilingüe 176

Apéndice D - Publicaciones **177**

Bibliografía **179**

Índice de figuras

1.1. Arquitectura de un sistema de traducción voz a voz (SST) 3
1.2. Tipos de sistemas de traducción automática 7

2.1. Sintetizador Klatt. 29
2.2. Ejemplo de contorno de entonación: *¿Cómo se llamaba el caballo de Calígula?* 34
2.3. Esquema del modelo de entonación de Fujisaki. 43
2.4. Parámetros Tilt. 45
2.5. Polinomios de Bézier de orden cuatro 46
2.6. Contorno de frecuencia fundamental aproximado usando curvas de Bézier con cinco coeficientes [Esc02b]. 46

3.1. Entrenamiento en dos pasos independientes: parametrización y entrenamiento del modelo. 60
3.2. Contorno de ejemplo que corresponde a la frase "Anda convulso el olimpo de las finanzas." 60
3.3. Inconsistencia debido al suavizado 61
3.4. Dos ejemplos de inconsistencia en la extracción de parámetros debido a requisitos de continuidad 62
3.5. Inconsistencia originada por el tipo de parametrización 63
3.6. Combinación de los pasos de modelado: entrenamiendo y parametrización conjuntos. 64
3.7. Ejemplo de datos de entrenamiento consistentes en dos oraciones. Las unidades prosódicas están numeradas del 1 al 5. 64
3.8. JEMA-Inicialización: Aproximación usando el contorno de la clase 0. 65
3.9. Ejemplo de complementariedad entre contornos. 66
3.10. JEMA-Partición: Mejor partición en la primera iteración. 66
3.11. JEMA-Optimización: Aproximación con dos clases en la primera iteración. . 67
3.12. Evolución del contorno JEMA. 73
3.13. Bucle de actualización de los parámetros de los comandos de acento y frase. 74
3.14. Distribución de la duración de la sílaba para diferente número de segmentos constituyentes 78

4.1. Contorno artificial correspondiente al modelo superposicional de Fujisaki. . 95
4.2. RMSE obtenido usando parametrización de Bezier para diferentes condiciones de ruido e información faltante en los datos de entrenamiento. 96

4.3. RMSE obtenido usando parametrización de Fujisaki para diferentes condiciones de ruido e información faltante en los datos de entrenamiento. 97
4.4. RMSE obtenido para los diversos modelos de entonación usando los datos de evaluación para el hablante femenino 99
4.5. Correlación obtenida para los diversos modelos de entonación usando los datos de evaluación para el hablante femenino 99
4.6. RMSE obtenido para los diversos modelos de entonación usando los datos de entrenamiento para el hablante masculino 100
4.7. Correlación obtenida para los diversos modelos de entonación usando los datos de evaluación para el hablante masculino 100
4.8. MOS de *naturalidad* obtenido para los diversos modelos de entonación . . . 102
4.9. MOS de *calidad* obtenido para los diversos modelos de entonación 102
4.10. RMSE obtenido para los diversos modelos de duración: hablante femenino . 105
4.11. RMSE obtenido para los diversos modelos de duración: hablante masculino 105
4.12. MOS de naturalidad obtenido para los diversos modelos de duración usando los datos de evaluación 107
4.13. MOS de calidad obtenido para los diversos modelos de duración usando los datos de evaluación 107

5.1. Esquema de generación de la prosodia utilizando la voz fuente. 119
5.2. Esquema de generación de la entonación utilizando la voz fuente. 122
5.3. Alineamiento usando grupos acentuales. 127
5.4. Ejemplo de alineamiento de grupos acentuales y asignación de clases. 129
5.5. Ciclo de mejora continua de clases. 130
5.6. Evolución de los parámetros de entrenamiento durante el agrupamiento. . . 132
5.7. Resultados experimentales usando el primer algoritmo propuesto en la dirección inglés → español. 136
5.8. Resultados experimentales usando el segundo algoritmo propuesto en la dirección inglés → español. Los datos poseen todos los signos de puntuación. 137
5.9. RMSE del logaritmo de la frecuencia fundamental usando el segundo algoritmo propuesto en la dirección catalán → español. Los datos solamente poseen puntos finales. 138
5.10. MOS de naturalidad obtenido para las diferentes condiciones experimentales usando los datos de evaluación. 140
5.11. Ejemplo de contornos predichos usando tanto información lingüística como la codificación del contorno de entrada. 141
5.12. Dispersión del ritmo en sílabas y acentos por segundo para el español e inglés británico 144
5.13. Correlación entre la velocidad de locución de los idiomas midiendo el ritmo a nivel de palabra usando el logaritmo de la duración de la misma 144
5.14. Correlación entre la duración de las palabras de los idiomas 144

5.15. Precisión de la sincronización utilizando algoritmos de compresión/expansión de pausas. 147

5.16. Precisión de la sincronización utilizando algoritmos de compresión/expansión de pausas y segmentos de voz. 148

A.1. Diagrama en bloques de los componentes del conversor texto a voz de la UPC: Ogmios. 163
B.1. Comparación de MSE entre diferentes imagenes del físico Eistein. De izquierda a derecha: original, disminución del contraste medio y contaminación con ruido gaussiano. [Imágenes extraídas del artículo de Wang(2009)] 171
B.2. Valores del coeficiente de correlacion Pearson para diferentes distribuciones de puntos. 171
B.3. Ejemplo de un *box-plot*. 173

Índice de tablas

3.1. Análisis de la correlación entre la duración de la sílaba y la duración segmental para cada fonema discriminado por el número de segmentos constituyentes de la sílaba. . . . 78
3.2. Entradas y salidas del transductor de estados finitos. $\overline{J}$ indica que no existe juntura terminal, y J indica que existe juntura terminal. . . . 87

4.1. MOS de naturalidad y calidad obtenido para los diversos modelos de entonación usando los datos de evaluación . . . 102
4.2. *Mann-Whitney-Wilcoxon test* de la evaluación subjetiva de la naturalidad de la entonación para los diversos modelos . . . 102
4.3. MOS de naturalidad y calidad obtenido para los diversos modelos de duración usando los datos de evaluación . . . 108
4.4. *Mann-Whitney-Wilcoxon test* de la evaluación subjetiva de la naturalidad de la duración para los diversos modelos . . . 108
4.5. Tabla de confusión . . . 109
4.6. Resultados experimentales del modelado de las junturas terminales del hablante femenino. . . . 110
4.7. Resultados experimentales del modelado de las junturas terminales del hablante masculino. . . . 110

5.1. Selección de los tiempos de referencia para la sincronización con el objeto de mantener una monotonía creciente. . . . 146
5.2. Selección de los tiempos de referencia para la sincronización usando pausas. 146
5.3. Resultados experimentales para los diferentes enfoques usando una comparación objetiva con una referencia. . . . 152
5.4. Resultados experimentales para los diferentes enfoques usando una comparación objetiva con una referencia, considerando que todas las pausas transferidas son correctas. . . . 152

A.1. Costos objetivos, donde B corresponde a valores binarios y C a valores continuos. . . . 167
A.2. Costos de concatenación, donde B corresponde a valores binarios y C a valores continuos. . . . 167

Capítulo 1

Introducción

Hoy en día existen en el mundo alrededor de setecientos idiomas. Cerca de quinientos están prácticamente extintos debido a múltiples razones, tales como el avance de las lenguas mayoritarias, guerras, colonialismo, políticas estatales en contra de las lenguas minoritarias, etc. El chino mandarín es el idioma con el mayor número de hablantes nativos, seguidos por el indio, español, inglés, árabe, portugués, bengalí, ruso, japonés y alemán. La existencia de diferentes idiomas y dialectos es una importante barrera para la comunicación humana. Las personas solamente pueden aprender una cantidad limitada de idiomas, siendo la principal tendencia el aprendizaje de aquellos que son de interés para el trabajo y viajes. Por estos motivos muchas lenguas minoritarias no son aprendidas, y son las mayoritarias las utilizadas para el diálogo entre interlocutores que no comparten una lengua materna común. Por ejemplo: un turista alemán y un conserje de hotel griego se comunican en inglés para poder comprenderse mutuamente.

Hay un pensamiento muy común en la gente: "...Yo hablo inglés, entonces no es necesario aprender otro idioma más...". Sin embargo, la última afirmación no es del todo cierta. De acuerdo al *CIA World Fact Book* solamente $5,6\,\%$ de la población total mundial habla inglés como lengua materna. Ese número se duplica cuando se cuenta aquellas personas que hablan inglés como segundo o tercer idioma. Por lo tanto, debido a que ni siquiera el inglés es un idioma ampliamente utilizado en el mundo, el desarrollo de dispositivos para la traducción automática voz a voz es una creciente necesidad que debe ser cubierta en el futuro próximo.

El área de la traducción voz a voz tiene como objetivo la traducción de la voz en un idioma y su reproducción en otro idioma en forma automática y sin la necesidad de intervención humana. Esto constituye un paso adelante con respecto a la traducción texto a texto, debido a que se realiza utilizando el habla, mediante la inclusión en el proceso de áreas tales como el reconocimiento automático de voz (ASR: Automatic Speech Recognition) y la generación de voz por computadora (TTS: Text-to-Speech Synthesis). El desarrollo de estas técnicas tendrá un impacto directo en muchas áreas [Wai08], tal como se menciona a continuación:

- *Instituciones internacionales.* Existen numerosas instituciones internacionales en las que cada día hay intervenciones en numerosos idiomas: Organización de las Naciones

Unidas, Parlamento Europeo, Organización Mundial del Comercio, para mencionar algunas de ellas. Allí se producen discusiones y discursos en varios idiomas usando tanto un estilo de habla leído como espontáneo. En estas reuniones los traductores humanos realizan su mejor esfuerzo para lograr la comprensión de lo expresado. Sin embargo, el estilo de traducción es mas bien neutral sin obedecer la forma de expresar las ideas de la persona que esta siendo traducida.

Un paso importante para la calidad de la traducción en este ámbito sería el uso de las tecnologías del habla (ASR, TTS), con el agregado de una prosodia rica en expresividad, en concordancia con la usada por el orador. Además, debido a que la traducción implica un alto costo, dado que cada vez son necesarios más idiomas debido a la expansión constante de estas instituciones, el uso de técnicas automáticas reduciría costos y permitiría la traducción a más idiomas. Este área de aplicación ha motivado a diversos proyectos de investigación, como es el caso de TC-STAR (Sección 1.3) [TCSTAR].

- *Difusión de noticias* Las agencias internacionales de noticias están distribuidas por todo el mundo y poseen audiencia en todos los idiomas. El uso de técnicas automáticas de traducción permitirá una difusión más rápida de las noticias reduciendo el tiempo de la tarea de traducción, y posiblemente aumentando la cobertura de idiomas. Por ejemplo, el proyecto europeo TC-STAR se abocó en una de sus tareas a la traducción de noticias, específicamente desde el chino mandarín al inglés europeo [TCSTAR, Tan02].

- *Reuniones.* Las tecnologías del habla serán también útiles en reuniones en las que se ven involucradas personas que hablan idiomas diferentes. Los sistemas proveerán traducciones voz a voz a los participantes con servicios adicionales, tales como archivado, resumen, comprensión del texto, acciones de apoyo (ilustraciones, movimientos de cámaras), etc [Füg06].

- *Educación.* La cooperación entre instituciones educacionales alrededor del mundo introduce el nuevo fenómeno de los estudiantes extranjeros en forma masiva. Es posible que ellos no posean un conocimiento suficiente del idioma cuando llegan al país que visitan. En tales situaciones las tecnologías del habla serían un importante apoyo, hasta lograr la adaptación del estudiante.

- *Turismo.* La aplicación de la traducción voz a voz en el ambiente del turismo será muy importante. Los turistas podrán hablar confortablemente en su lengua materna cuando se encuentran en otro país, permitiendo un mejor aprovechamiento de su tiempo, de las experiencias y de los lugares visitados [Cet99][Bur03].

- *Asistencia médica.* Los extranjeros pueden tener serios problemas de comunicación cuando necesitan asistencia medica en países donde no se habla su idioma. Las tecnologías del habla pueden ofrecer un importante servicio en tales situaciones para la comunicación entre el especialista y el paciente [Wai03][Bur03][Bou08].

En las siguiente sección (Sección 1.1) se introducirán nociones del área de la traducción voz a voz, explicando las tecnologías que se ven envueltas, e indicando sus problemas y limitaciones.

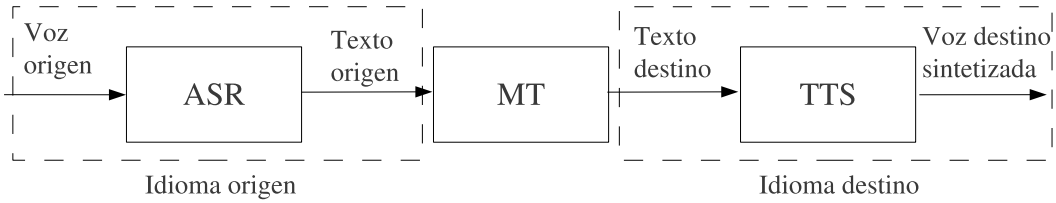


Figura 1.1: Arquitectura de un sistema de traducción voz a voz (SST)

En la Sección 1.2 se dará una reseña de los proyectos desarrollados en el área, lo que permitirá ver la variedad de enfoques propuestos con sus fortalezas y debilidades. Se dedica especial atención al proyecto TC-STAR en la Sección 1.3, ya que fue donde se enmarcó esta tesis.

Finalmente, los objetivos de la tesis se detallan en la Sección 1.4. Allí se indican los diferentes modelos prosódicos que se describirán en el resto de la tesis, con las condiciones de entrenamiento y evaluación.

1.1. Arquitectura de un sistema de traducción voz a voz

La traducción voz a voz es un interesante campo en el que muchas empresas e instituciones están dedicando esfuerzos para lograr avances en la aplicabilidad de la técnica en los distintos dominios mencionados en la sección anterior.

En la Figura 1.1 se puede observar que un sistema de traducción voz a voz (SST: Speech-to-Speech Translation) se puede dividir en tres componentes básicos bien diferenciados [Ekl95]:

- *Reconocimiento automático del habla (ASR: Automatic Speech Recognition).* La entrada de un SST es la voz del locutor que se desea traducir. El sistema de ASR convierte la voz origen en texto usando técnicas estadísticas de modelado acústico y decodificación.
- *Traducción automática (MT: Machine Translation).* El texto en el idioma origen se traduce al idioma destino usando un sistema de traducción automática. En esta etapa se tienen en cuenta ciertas particularidades de los idiomas origen y destino, tales como el ordenamiento diferente de las palabras en la oración, declinaciones, etc.
- *Síntesis de voz (TTS: Text-to-Speech).* El texto traducido es la entrada del sistema de síntesis de voz. Este sistema convierte el texto en voz usando una prosodia generada automáticamente a partir del texto.

Para tener una idea más detallada del funcionamiento de cada uno de los componentes de un sistema de traducción voz a voz, daremos en las siguientes secciones una introducción al reconocimiento automático del habla (Sección 1.1.1), traducción automática

(Sección 1.1.2) y conversión texto a voz (Sección 1.1.3). Esta información resultará útil para la comprensión de la descripción de los diferentes proyectos en las secciones 1.2 y 1.3. Además, permitirá entender mejor la propuesta de esta tesis de la Sección 1.4 y las condiciones experimentales sobre las cuales se ha trabajado con el objetivo de desarrollar mejores modelos para la generación de prosodia en el campo de la traducción voz a voz.

1.1.1. Reconocimiento automático del habla

En general, las personas usan una gran variedad de recursos en el momento de expresarse verbalmente: palabras, entonación, intensidad, variaciones en el ritmo o la duración de los fonemas, pausas, diferentes tipos de fonación, etc.

La tarea de un sistema de reconocimiento automático del habla consiste en obtener una secuencia de palabras (o etiquetas) que son la representación textual de una señal acústica. Dicha tarea no es trivial, ya que en muchas circunstancias la señal acústica no contiene solamente palabras y pausas, sino también disfluencias, sonidos ambientales, ruidos de origen humano (labios, respiración), entre otros. Como resultado de esto, en algunos casos es posible que el reconocimiento de voz también involucre la detección de estos sonidos para evitar una confusión en el intento de reconocerlos como palabras [Tem06].

Para realizar esta tarea, los sistemas de reconocimiento automático del habla hacen uso de una serie de herramientas de modelado estadístico y decodificación. En este enfoque se aplican un conjunto de simplificaciones para obtener una solución implementable tanto desde el punto de vista de la precisión estadística, como de la complejidad y tiempo de procesamiento para la decodificación.

Una de las primeras cosas que se asume es que el vocabulario a reconocer estará limitado. Su tamaño puede ir de unas pocas palabras (por ejemplo: sistemas de reconocimiento de órdenes verbales) a decenas de miles de palabras (sistemas de reconocimiento de gran vocabulario). No es posible para un ASR reconocer palabras desconocidas debido a que no le sería posible encontrar las fronteras de las mismas. Por ejemplo, la oración *"la casa de la pradera está habitada por un ermitaño"* se podría pronunciar como *"lacasadelapradera estáhabitada porunermitaño"*. Como se puede observar, existen pausas después de las palabras *pradera* y *habitada*. Sin embargo, el resto de las palabras son pronunciadas sin dar ningún indicio del fin de una y del comienzo de la otra. Por lo tanto, es necesario conocer las palabras para poder encontrar sus fronteras.

Los sistemas de reconocimiento automático del habla que se desarrollan hoy en día asumen que es posible realizar dicha tarea basándose en un modelado estadístico de la señal acústica (modelado de la generación acústica de las palabras del idioma) y el lenguaje (modelado de la construcción del discurso del idioma utilizando palabras). Para ello se usan datos de entrenamiento (señal acústica, y transcripción ortográfica y fonética de la misma) para obtener los parámetros de los modelos estadísticos. La cantidad de datos de entrenamiento está acotada por el volumen del corpus disponible, y por ello es necesario usar modelos cuyos parámetros puedan ser obtenidos en forma confiable teniendo en cuenta esta limitación.

Los humanos usamos una gran cantidad de fuentes de información para entender

el habla: información acústica, gestos, gramática, semántica, contexto, etc. Todas estas fuentes complementarias a la información acústica permiten que no sea necesario escuchar la secuencia completa de fonemas (debido a fonemas mal pronunciados, palabras incompletas, ruidos en la señal, etc.) para reconocer la secuencia de palabras. Este comportamiento ha conducido al uso en los ASR del modelado estadístico del lenguaje (o modelo del lenguaje) para estimar la probabilidad de las diferentes posibles secuencias de palabras del lenguaje.

Por otra parte, las palabras están compuestas por una secuencia de fonemas. Los ASR realizan un modelado acústico estadístico de los fonemas (usando semifonemas, fonemas, trifonemas, etc) para obtener modelos estadísticos con parámetros mejor estimados que aquellos que se pueden obtener usando otras unidades más grandes, tales como sílabas, palabras u oraciones. Los parámetros de palabras y oraciones son más difíciles de estimar debido a la cantidad limitada de ellas (número de repeticiones) dentro de los datos de entrenamiento.

Por lo tanto, en general se puede decir que los ASR asumen que la tarea de reconocimiento automático del habla puede ser dividida en dos partes: un modelado estadístico del lenguaje y un modelado acústico de los fonemas.

Estas simplificaciones introducen limitaciones y dificultades en los ASR, tal como se describe en la siguiente sección.

Dificultades y limitaciones de los ASR

Los sistemas de reconocimiento automático del habla enfrentan gran variedad de problemas. Algunos de ellos están lejos de ser solucionados debido a las limitaciones de las técnicas actuales [For03]:

- **Capacidad de comprensión de los humanos.**

  Como se mencionó anteriormente, los humanos usamos más fuentes de información que aquella que percibimos a través de nuestros oídos para entender lo que estamos escuchando. En este proceso se usan conocimientos acerca del hablante, el contexto, el ambiente, entre otros. Todo esto contribuye a solucionar problemas de ambigüedad que aparecen en el proceso de entendimiento, tales como homófonos y la ausencia de indicios acústicos de las fronteras entre palabras. Los ASR carecen de tal saber, y por lo tanto sus prestaciones son más pobres que las obtenidas por humanos.

- **Lenguaje hablado y lenguaje escrito.**

  El lenguaje hablado es esencialmente diferente del lenguaje escrito. En general, la comunicación escrita es un proceso en un solo sentido, mientras que el habla está orientado al diálogo: existe una realimentación a nuestro interlocutor, se negocia el significado de las palabras, nos adaptamos al oyente, etc. Un ejemplo de esto es la gramática del habla espontánea, la cual está lejos de ser formal. Las repeticiones y los borrados introducen una gran cantidad de palabras incompletas que se convierten en palabras fuera del vocabulario para un ASR [Sri06].

- **Factores relacionados con el hablante.**

  La variabilidad intrahablante e interhablante es otra fuente de dificultades para los ASR. La realización del habla es un proceso estocástico. Si las mismas palabras son pronunciadas repetidas veces por un hablante, la señal resultante nunca será igual.

  Por otra parte, existen factores adicionales que contribuyen a esta variabilidad, tales como el sexo, anatomía, factores sociales y geográficos, etc. En algunas situaciones las diferencias entre dialectos es tan grande que son considerados diferentes idiomas para los propósitos de los ASR o bien son modelados con diferentes inventarios de fonemas [Cab04].

- **Factores ambientales.**

  El habla se produce en un ambiente que es propenso a la interferencia por parte de muchos otros sonidos, tales como los ventiladores de los ordenadores, ruidos de manipulación de muebles, motores de automóviles, apertura y cierre de puertas, otros hablantes, etc. Todos estos sonidos son considerados como ruido por ser información no deseada en la señal, ya que perturban la claridad de la señal del habla. Para contrarrestar estos factores los ASR hacen uso de diversas técnicas : identificación de ruidos, filtrado, ecualización del canal, etc.

Una importante limitación de los ASR es que se enfocan principalmente en el reconocimiento de palabras, y en algunos casos también *fillers* y diversos tipos de metadatos (puntuación, disfluencias, atributos del hablante, etc) [Liu04b]. Mucha información prosódica (entonación, duración segmental, ritmo) no se encuentra en la salida entregada por el ASR, cuando podría resultar de utilidad para mejorar tanto la traducción automática como la conversión texto a voz en un traductor voz a voz. Además, otras tareas se podrían ver beneficiadas con dicha información, como es el caso de la sincronización audio/video y el subtitulado automático.

1.1.2. Traducción automática

La traducción automática consiste en traducir un lenguaje natural en otro usando un algoritmo en una computadora. Este problema ha sido particularmente difícil de resolver en el área de la inteligencia artificial desde hace varias décadas. Los enfoques iniciales fallaban debido a que muchos fenómenos complejos convierten a la traducción automática en un problema inmanejable.

Hace pocos años hubo grandes mejoras en el campo de la traducción automática a través del uso de enfoques estadísticos. Para entrenar sistemas de traducción automática estadísticos se usan grandes volúmenes de datos del orden de los gigabytes de texto (o terabytes, como es el caso de Google [Bra07]).

La arquitectura de un sistema de traducción automático se puede organizar en una estructura piramidal compuesta de tres niveles: directo, transferencia e interlingua. La profundidad del análisis del sistema está directamente relacionado con el nivel. La parte inferior de la pirámide es la forma más primitiva de traducción: reemplazo de cada palabra en un idioma por una palabra en el otro idioma sin tener en cuenta aspectos tales como

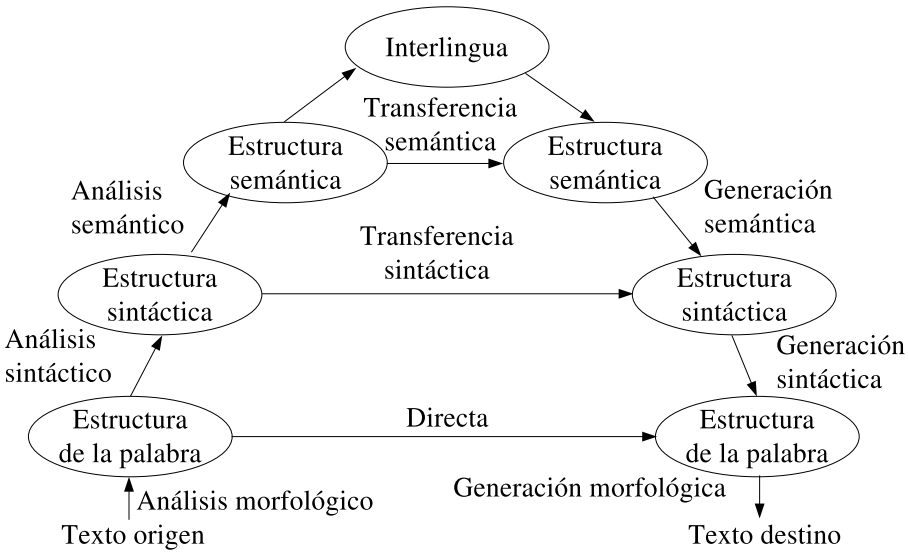


Figura 1.2: Tipos de sistemas de traducción automática

expresiones idiomáticas y diferencias gramaticales entre los idiomas (aspectos considerados en enfoques modernos de traducción estadística que se enfocan en la palabra). En la parte superior de la pirámide se encuentra la traducción usando un idioma abstracto: el enfoque interlingua [Dor98].

Para poder entender un poco mejor cada nivel, haremos una breve descripción de cada uno de ellos:

- *Transferencia directa.* La arquitectura de traducción directa consiste en generar la traducción en el idioma destino reemplazando directamente las palabras del idioma origen. Sin embargo, el orden de las palabras en el idioma destino puede ser diferente que el del idioma origen. Para considerar este aspecto, en la actualidad se aplican técnicas de reordenamiento bilingüe conjuntamente con modelos de lenguaje de los idiomas que tienen un gran éxito, como es el caso del traductor de Google [Bra07].

- *Transferencia sintáctica.* El enfoque inicial de los sistemas con arquitectura de transferencia fue el uso de la información en el nivel sintáctico. Se usan diversas representaciones del idioma origen para transformarlas en una representación sintáctica adecuada en el idioma destino. La traducción resultante tiene una estructura apropiada desde el punto de vista sintáctico, lo cual es un paso adelante con respecto a los sistemas con arquitectura directa. La principal limitación de estos sistemas es que el análisis sintáctico automático es un problema sin resolver, principalmente en oraciones complejas con muchas oraciones subordinadas y frases nominales, verbales y preposicionales.

- *Transferencia semántica.* Teóricamente, la arquitectura de transferencia semántica brindaría los mejores resultados de traducción. Un análisis del contexto (discurso y pragmático) es necesario conjuntamente a un análisis semántico profundo. Debido

a la complejidad de la tarea, este enfoque se puede aplicar solamente en dominios pequeños donde el análisis, las reglas y el vocabulario bilingüe permiten cubrir todo el dominio.

- *Interlingua.* La principal idea detrás del enfoque de interlingua es traducir el idioma origen en una lengua abstracta. En el caso de que la tarea se pueda hacer exitosamente, solo será necesario desarrollar un sistema de traducción automática para cada idioma y el lenguaje intermedio (interlingua). La traducción de un idioma a otro se hace pasando a través del lenguaje abstracto. El enfoque de interlingua es muy interesante porque reduce el tiempo de desarrollo para cubrir nuevos idiomas. Sin embargo, el análisis automático del texto es todavía un campo en desarrollo, y por ello interlingua puede ser aplicado solamente a dominios muy restringidos, tales como reservas (aviones, buses, trenes, hoteles) o consulta de bases de datos, para mencionar algunos de ellos.

Dificultades de los sistemas de MT

Los desafíos en el área de la traducción automática se pueden examinar usando dos dimensiones: consideraciones lingüísticas en lo referido al orden sintáctico de las palabras y a la ambigüedad semántica, y consideraciones operacionales, tales como extensibilidad, mantenibilidad, interfaz con el usuario, etc [Dor98].

Dentro de las consideraciones lingüísticas se encuentran:

- **Ambigüedad sintáctica**. Una secuencia de sintagmas preposicionales y nominales coordinados pueden inducir a ambigüedades. Por ejemplo: *"Observé al hombre en la montaña con un telescopio"*. En esta última oración no es posible decidir quién posee el telescopio debido a la falta de información contextual.
- **Ambigüedad léxica**. Una palabra con dos significados posibles en el idioma origen puede ser difícil de traducir en aquellos casos donde el contexto sintáctico disponible sea insuficiente. Por ejemplo: *book* se puede traducir como libro o como reservar.
- **Ambigüedad semántica**. Esto corresponde al caso de una palabra con significados diferentes y con la misma función en la oración. Por ejemplo, homógrafos tales como *ball: pelota, baile* o casos de polisemia *kill: matar, terminar.*
- **Ambigüedad contextual**. Un ejemplo de este tipo de ambigüedad se puede observar en la siguiente oración: *"The computer outputs the data; it is fast."* En este caso solamente se puede saber que *fast* alude a *computer* porque rápido no es un atributo posible para *data.*
- **Selección léxica**. Selección de las palabras apropiadas que transporten el significado adecuado de la oración del idioma origen. Por ejemplo, la palabra *fish* en inglés se puede traducir como *pez* o *pescado*, y es en español donde se diferencia si el animal está en el agua o no.

- **Divergencia estructural**. La estructura sintáctica puede ser diferente en el idioma origen y destino. Por ejemplo: *"John entered the house → Juan entró en la casa"*. El sintagma nominal *the house* en inglés se traduce como el sintagma preposicional *en la casa* en español.

Las consideraciones lingüísticas no son los únicos desafíos de los sistemas de traducción automática. Los desafíos operacionales surgen en el momento de la implementación del sistema: alineamiento de corpus bilingües, adaptación al dominio y preparación de los datos de entrenamiento.

Una parte importante del campo de la traducción estadística es el alineamiento automático de las palabras en un corpus bilingüe. Esta tarea es casi imposible de hacer manualmente debido al volumen de datos necesarios para el entrenamiento del sistema. La complejidad de esta tarea puede ser analizada a través de un ejemplo: si el sistema analiza la traducción *I stabbed John→Yo le di puñaladas a John*, el alineamiento se puede hacer apropiadamente manualmente: *I→Yo*, *stabbed→le di puñaladas a* y *John→John*. Sin embargo, dependiendo de la cantidad de información disponible para el alineamiento estadístico, es posible que el sistema tenga dificultades para realizar tal alineamiento en el caso de la estructura compleja *stabbed→le di puñaladas a*.

La selección del dominio del sistema es otro punto clave también, ya que restringirá el léxico y la gramática, reduciendo los problemas de la ambigüedad léxica, homógrafos, polisemia, metonimia, ambigüedad contextual, selección léxica y la generación de tiempos verbales.

Por otra parte, la cantidad de texto necesario para cubrir las demandas de un dominio dado varía desde los pocos megabytes hasta varios gigabytes. Dichos corpus deben ser analizados cuidadosamente porque es común encontrar problemas que pueden perjudicar al sistema de alineamiento automático, tales como diferencias en el número de oraciones del idioma origen y destino, cambios en el orden de las oraciones en un párrafo, problemas de formato, etc. Los algoritmos de alineamiento de texto bilingües sufren gravemente en tales situaciones.

1.1.3. Conversión texto a voz

En el siglo XVIII fue cuando los científicos construyeron los primeros modelos mecánicos capaces de producir vocales. Los esfuerzos primigenios que producían voz sintética fueron hechos hace doscientos años por Kratzenstein y Wolfgang von Kempelen [Fla72, Fla73, Sch93].

En 1779, en San Petersburgo, el profesor ruso Christian Kratzenstein explicó las diferencias fisiológicas entre las cinco vocales (/a/, /e/, /i/, /o/, y /u/), y construyó dispositivos mecánicos capaces de producirlas artificialmente mediante resonadores acústicos similares al tracto vocal. Los resonadores eran activados por lengüetas vibrantes (como en los instrumentos musicales). Por ejemplo, el sonido /i/ era producido soplando en un tubo más bajo sin lengüeta causando un sonido similar a una flauta.

Unos pocos años después, en el año 1791 en Viena, Wolfgang von Kempelen introdujo

su “Máquina de Hablar Acústico-Mecánica”, la cual era capaz de producir sonidos simples y algunas combinaciones de sonidos [Kla87, Sch93].

Sin embargo, recién en las últimas décadas del siglo pasado ha habido un gran desarrollo en el campo de la síntesis de voz debido a la invención de la computadora. Este dispositivo programable permitió el uso de nuevos paradigmas en la generación de voz, tales como síntesis articulatoria, síntesis por formantes, síntesis por concatenación, aprendizaje basado en datos, uso de algoritmos de programación dinámica, entre otros, que provocaron un salto significativo en la calidad de la voz generada.

Como consecuencia del advenimiento de las computadoras, los sintetizadores de voz han derivado en sistemas de conversión texto a voz. Dicho texto digital de entrada puede poseer una gran variedad de formatos, tales como texto plano, texto con formato (libros, revistas), e-mail, SMS, texto escaneado, etc.

La calidad de voz que alcanzan hoy en día es difícil de distinguir del producido por un humano en frases cortas, pero todavía se notan diferencias cuando se sintetizan párrafos y textos extensos. La continua evolución del área ha producido la continua aparición de nuevas aplicaciones de esta tecnología.

Probablemente una de las aplicaciones más importantes desde el punto de vista social es la ayuda en la lectura y comunicación de personas con discapacidades. El uso de audio es más efectivo que otros sistemas, tales como los caracteres Braille, para la comprensión de textos. Antes de la existencia de esta tecnología, como es el caso de “The Intel Reader”, muchos libros en audio era grabados por humanos usando cintas de audio. Claramente, la realización de tal grabación es cara y emplea considerable tiempo, por lo que el número de audiolibros es sólo una ínfima parte del material publicado.

Otro ejemplo de la utilidad de esta tecnología es su aplicación a la comunicación de las personas sordas, o bien que no son capaces de hablar apropiadamente. Los sintetizadores de voz dan una oportunidad de comunicarse de una manera más clara con personas que no entienden el lenguaje de señas o les resulta dificultoso entender el habla de una persona discapacitada. Un caso conocido de esta aplicación es el sistema de conversión texto a voz Neospeech usado por el físico teórico Stephen Hawking.

Además de las aplicaciones que permiten mejorar la calidad de vida de los discapacitados, los sintetizadores de voz se usan en ocasiones para dar información. Tal es el caso de los sistemas de navegación para automóviles, en donde pueden dar un gran número de indicaciones mencionando incluso el nombre de las calles.

Como ya hemos mencionado al inicio del capítulo, otro importante campo de aplicación de los sintetizadores de voz es la traducción automática voz a voz. Resulta preferible la utilización de dispositivos de traducción automáticos voz a voz móviles que se puedan llevar a cualquier parte en lugar de un traductor humano que no estará disponible en todas las situaciones: reuniones, registración, restaurantes, etc. Los dispositivos móviles son baratos y cada vez poseen mejores servicios que permiten su aplicación a la traducción.

En un conversor texto a voz (TTS: Text-to-Speech) intervienen varios módulos que realizan sucesivas transformaciones del texto de entrada e incorporan una gran cantidad de información adicional útil para generar el sonido sintetizado, tales como el procesamiento

del texto y la generación de la prosodia, para luego, finalmente, producir la síntesis de la voz. En general los componentes de cada módulo son:

- *Procesamiento del texto.* Uno de los primeros pasos es la **normalización del texto** para expandir abreviaciones, convertir números y fechas en texto, etc. Esta es una tarea importante ya que es necesario para el correcto funcionamiento de los siguientes módulos que el texto este constituido solamente por grafemas. En ocasiones se realiza también un etiquetado morfosintáctico, que resulta necesario para tareas como la conversión de grafemas a fonemas y la generación de junturas terminales en la prosodia.

  A continuación se realiza la **conversión de grafemas a fonemas** y la *silabificación* para obtener la secuencia de fonemas necesaria para reproducir el texto, incluyendo la información sobre sílabas que también es importante para los siguientes módulos.

- *Prosodia.* En este componente se genera la información prosódica para poder producir la voz. Para ello se predicen las frases entonativas y la entonación de la oración, y también la duración y la energía de los fonemas, etc. La correcta generación de esta información repercutirá directamente en la naturalidad y expresividad del sistema.

  La importancia de la prosodia se ve reflejado en que puede modificar el contenido de un mensaje. Por ejemplo, la frase *irás al cine* es una afirmación si existe una declinación de la entonación al final de la oración. Sin embargo, será interpretada como una pregunta si la frecuencia fundamental sube al final de la frase.

- *Generación de la voz.* En este último componente es donde se genera la voz considerando la información provista por los módulos de **Procesamiento del texto** y **Prosodia**, la cual está constituida por la secuencia de fonemas con la información relativa a su duración, energía, frecuencia fundamental y posición dentro de la frase entonativa, la oración y el discurso.

  La generación se puede realizar usando diversos enfoques, tales como selección de segmentos paramétrica-estadística, síntesis por formantes o síntesis articulatoria. Los mismos serán explicados en más detalle en la Sección 2.1.3.

Desafíos de la conversión texto a voz

Existe una serie de dificultades que tienen que afrontar cada uno de los módulos mencionados anteriormente:

- **Procesamiento del texto.** Desde la conversión de grafemas a fonemas de palabras fuera del vocabulario o con múltiples transcripciones hasta el tratamiento de diferentes formas de representación de números, fechas, abreviaturas; todas estas constituyen complicadas tareas que distan en muchos casos de estar resueltas.

- **Prosodia.** La generación de la prosodia es una tarea que envuelve un conocimiento y unas habilidades que están lejos de poder ser realizadas por los ordenadores. Los mismos no poseen información sobre el mundo, la actualidad, ni tampoco tienen una

opinión. Por lo tanto, la prosodia generada en algunas ocasiones será inadecuada para el contenido del texto o bien demasiado neutra para ser agradable y natural. Un desafío creciente lo representa la generación de voz variada, emocional, con variación de ritmo, tal como lo indicaron en su momento Bailly et al. [Bai03], acerca del habla expresiva.

- **Generación de la voz.** Las limitaciones de las técnicas para generación de la voz provocan la aparición de fenómenos tales como ruidos de coarticulación, trayectorias de formantes inadecuadas, etc [Kla01].

Otros desafíos observados por Simon King [Kin10] en la conversión texto a voz es el desarrollo de aplicaciones en condiciones que presentan ciertas restricciones. Entre ellos se encuentra el desarrollo de un conversor texto a voz con muestras de un paciente con desórdenes del habla, donde la cantidad y la calidad de la voz presentan grandes limitaciones.

Un desafío que se encuentra en el ámbito del proyecto marco de esta tesis es la síntesis de voz en un idioma objetivo que suene como un hablante en particular, cuando solamente se dispone de muestras de voz de dicho hablante en otro idioma. El abordaje de esta problemática se vuelve aún más compleja cuando no se dispone conocimiento detallado de la fonología u otros recursos, tales como diccionarios de pronunciación, modelos prosódicos, o habla grabada sin transcripción ortográfica.

1.2. Proyectos relacionados con la traducción voz a voz

Muchos proyectos han estado y están centrados en la traducción voz a voz. Excepto los proyectos más recientes, todos ellos se han centrado en dominios restringidos tales como inscripción en conferencias y hoteles, reservas de viajes en aviones y trenes, alquiler de autos, etc. Esto se debe a que un sistema de dominio sin restricciones esta lejos de ser realizable debido a su complejidad y al grado de avance de las técnicas involucradas: ASR, MT y TTS.

A continuación se hará una breve reseña de una variedad de proyectos que han involucrado a empresas y universidades de todo el mundo: JANUS, C-STAR, Verbmobil, Nespole!, MASTOR, DIPLOMAT/Tongues, ATR-MATRIX y TC-STAR. Muchos de estos proyectos utilizan la modalidad oral en la entrada, pero no en la salida.

JANUS fue uno de los primeros sistemas diseñados para la traducción automática del habla, y fue desarrollado desde fines de los años ochenta por los Laboratorios de Sistemas Interactivos de la Carnegie Mellon University y la Universität Karlsruhe, y desde principios de los años noventa también aportaron a su desarrollo ATR (Japón) y Siemens AG (Alemania). Desde entonces fue extendido a tareas más avanzadas, contribuyendo a varios esfuerzos en el área, tales como Enthusiast (EEUU) y Verbmobil (Alemania).

La primera versión de JANUS (JANUS-I) [Wos93] solamente procesaba habla sintácticamente correcta (habla leída) usando un pequeño vocabulario de 500 palabras. A pesar de que JANUS-II y JANUS-III [Lav96] (este último una versión mejorada) operaban todavía en dominios limitados, tales como planificación de viajes, estos podían manejar habla

espontánea en diálogos con vocabularios de más de 10.000 palabras, trabajando con idiomas tales como inglés, alemán, japonés o coreano, usando el enfoque interlingua. Gracias a ampliaciones desarrolladas en IRST y CLIPS++, también se podía trabajar con italiano y francés. La síntesis de voz era provista por dispositivos comerciales, tales como el sistema Digital DECtalk DTCO1 para el alemán, o el Panasonic Text-to-Speech System EV-3 para japonés. Cada uno de estos sistemas utilizaba una representación textual o fonética, y producía los sonidos de la oración a través de un parlante.

Otro ejemplo de aplicación de interlingua en la traducción automática es el Consorcio para la Investigación Avanzada de la Traducción del Habla (Consortium for Speech Translation Advanced Research: C-STAR), que en un principio surgió a causa de colaboraciones informales bilaterales entre laboratorios interesados en la investigación sobre traducción automática del habla. Al comienzo, en 1991, estaba constituido por sus miembros fundadores: ATR Interpreting Telephony Laboratories (Kyoto, Japon), Carnegie Mellon University (Pittsburgh, EEUU), University of Karlsruhe (Karlsruhe, Alemania) y Siemens AG (Munich, Alemania). Este consorcio demostró en 1993 por primera vez en público que la traducción automática del habla usando enlaces internacionales de comunicación era posible [Yam95].

Desde entonces el consorcio continuó en 1993 como C-STAR II, comenzando la segunda fase de la investigación. Debido a la importancia que cobró la demostración, en esta fase el consorcio creció sustancialmente incluyendo 6 miembros de 6 países (Japon, Corea, EEUU, Alemania, Italia y Francia) y 14 miembros afiliados en 10 países. Los esfuerzos se concentraron en la traducción de habla espontánea con prototipos que podían aceptar vocabularios de 10.000 a 100.000 palabras. A causa de la variedad de países participantes, era posible traducir diálogos en 6 idiomas. C-STAR III incorporó el chino, y el área de aplicación fue el dominio turístico usando una línea telefónica. Luego del reconocimiento, el análisis y la generación del lenguaje, el texto de salida se transmitía a cada uno de los miembros para que cada uno de ellos hiciera la síntesis con el sistema elegido [Wai96].

Un importante proyecto en la historia de la traducción voz a voz fue Verbmobil, debido al gran número de descubrimientos científicos en las áreas del procesamiento del habla, lenguaje y discurso, traducción de diálogos, generación del lenguaje y síntesis de voz, lo cual ha sido documentado en más de 800 publicaciones y un libro [Wah00].

Verbmobil[Wah00] fue un proyecto internacional en el que estuvieron envueltas diversas compañías y universidades, tales como Alcatel (Stuttgart), Daimler-Benz AG (Stuttgart), IBM (Stuttgart), Philips GmbH (Aachen), Siemens Aktiengesellschaft (Berlin), Daimler-Benz Aerospace AG (Ulm), Carnegie Mellon University (EEUU), University of Stuttgart, University of Bonn, entre otras. El primer sistema completamente integrado, el demostrador Verbmobil, fue presentado en público durante CeBIT (CeBIT: Centrum für Büro und Organisationstechnik) 1995. Ese demostrador reconocía oraciones en alemán en el contexto de la negociación de citas usando un vocabulario de 1292 palabras. El sistema era capaz de analizarlas y las traducía al inglés. El prototipo presentado en CeBIT 1997 tenía un vocabulario ampliado de 2500 palabras y reconocía japonés usando 400 palabras, traduciéndolas también al inglés.

Durante la primera fase (1993-1996) se desarrolló un sistema de traducción voz a voz de habla espontánea independiente del hablante para conversaciones relacionadas con la

concertación de citas, planificación de viajes y reservas de hotel. Esta elección implicó un rediseño del sistema de reconocimiento automático del habla para tolerar disfluencias y otros fenómenos del habla espontánea.

En contraste con el lenguaje escrito, el lenguaje hablado no tiene información sobre puntuación. Como consecuencia de esto, en este proyecto se hizo uso de un análisis del énfasis utilizado en el habla, y también de la división del discurso en frases, para solventar esa ausencia de información. El módulo de prosodia reconocía fronteras de oración y frases haciendo uso de información sobre junturas terminales, entonación, duración y energía de la voz. Estos datos eran provistos al módulo de análisis sintáctico con las fronteras de cada oración.

En el proyecto se aplicaron análisis sintácticos y semánticos simultáneos para examinar los resultados del ASR en la búsqueda de frases gramaticalmente correctas. De esta manera, el planificador de discursos agrupaba los actos de diálogo en fases con capacidad para reproducir la discusión en forma abreviada.

El módulo de procesamiento sintáctico-semántico entregaba los datos analizados al módulo de transferencia, y era este último quien realizaba una representación abstracta en el lenguaje destino, para que luego el generador convirtiera los predicados semánticos en oraciones sintácticamente correctas. El módulo final del sistema era un TTS, el cual reproducía la traducción al inglés en una forma tan natural y comprensible como sea posible, y con el énfasis requerido.

Luego de varios años de investigación y desarrollo Verbmobil alcanzó importantes metas, tales como una tasa de reconocimiento de palabras cercano al 75 % para habla espontánea, con alrededor del 80 % de las traducciones en forma correcta preservando la intención del hablante, o una tasa de éxito del 90 % para tareas de diálogo con usuarios reales.

Otro proyecto fue el denominado ATR-MATRIX. ATR (Advanced Telecommunications Research Institute International) comenzó sus estudios sobre traducción voz a voz a mediados de los años ochenta, y desarrolló el sistema de traducción del habla multilenguaje llamado ATR-MATRIX (ATR's Multilingual Automatic Translation System for Information Exchange) [Tak98, Sum99].

El sistema poseía un sistema de reconocimiento automático del habla con una alta precisión para habla espontánea, que se encuentra descrito junto al sistema de síntesis de voz basado en corpus en Takezawa et al. [Tak99]. La traducción automática se realizaba usando un enfoque basado en ejemplos, llamado TDMT (Transfer-Driven Machine Translation) [Fur95].

Las características principales de ATR-MATRIX eran:

- La traducción se realizaba entre los idiomas inglés y japonés, en ambas direcciones.
- El reconocimiento automático del habla, la traducción automática y la síntesis de voz podían correr en tiempo real en una computadora personal.
- El sistema era manos libres, permitiendo comenzar y parar de hablar en una modalidad de conversación completamente full-duplex.

- El sistema fue desarrollado para funcionar con habla natural, con un tratamiento adecuado de interjecciones y expresiones informales.
- La salida de voz posee el mismo sexo que la voz de entrada.

Otros proyectos en el área de la traducción voz a voz se enfocaron en otorgar buenas prestaciones con bajos requerimientos de hardware: NESPOLE!, MASTOR y DIPLOMAT/TONGUES.

El objetivo de NESPOLE! era proporcionar un sistema de traducción automática de voz aplicable a las necesidades en el área de e-commerce y e-service. El proyecto estuvo constituido por tres grupos de investigación europeos (IRST, Trento, Italia; ISL, Universität Karlsruhe, Alemania; y CLIPS, Université Joseph Fourier, Grenoble, Francia) y un grupo estadounidense (ISL, Carnegie Mellon University, Pittsburgh, EEUU). También lo integraron dos participantes industriales: APT (Trento, Italia) y AETHRA Telecomunicazioni (Ancona, Italia). El soporte financiero estuvo dado por la Comisión Europea y el US NSF.

La arquitectura de NESPOLE! [Lav01] estaba basada en la filosofía cliente-servidor, usando paquetes de software como NetMeeting. A través del componente que en el proyecto se denominó mediador, se regulaba el canal de comunicación entre los integrantes de la misma y también se establecían los enlaces con los servidores de traducción automática (Human Language Technology servers). Estos últimos son los que proporcionaban los servicios de reconocimiento de voz y traducción entre idiomas. El modulo de generación producía un texto en el lenguaje destino en base a la información de la representación interlingua. Dicho texto se enviaba al módulo de síntesis de voz, como es el caso del sintetizador Euler TTS para el francés [Bes01].

Otro sistema desarrollado con bajos requerimientos de hardware fue MASTOR. La investigación en este proyecto se inició en el año 2001 como un proyecto de IBM, el cual fue seleccionado por DARPA CAST para otorgarle fondos para su desarrollo.

MASTOR [Liu03] combinaba varias tecnologías desarrolladas por IBM en las áreas del reconocimiento automático del habla, entendimiento del lenguaje natural y la síntesis de voz. El acoplamiento del ASR con el componente de entendimiento del lenguaje natural permitía obtener un sistema robusto que mitigaba los efectos de los errores de reconocimiento y la gramática incorrecta del habla coloquial.

El sistema desarrollado permitía traducir en forma bidireccional inglés y mandarín con un gran vocabulario de más de 30.000 palabras en varios dominios, tales como viajes, diagnóstico médico de emergencia y tareas relacionadas con fuerzas de defensa y seguridad.

Finalmente, describiremos el proyecto DIPLOMAT [Fre97]. El mismo fue diseñado para explorar la posibilidad de crear rápidamente sistemas de traducción voz a voz bidireccionales. El objetivo era la generación de un sistema que pueda traducir entre un idioma nuevo e inglés en cuestión de días o semanas, con posibilidad de mejorar la calidad al cabo de unos meses.

El sistema de entendimiento se basaba en SPHINX II [XH92][Rav96], aplicando técnicas que permitían desarrollar rápidamente modelos acústicos y de lenguaje [Rud95]. La

traducción automática se realizaba usando MEMT (Multi-Engine Machine Translation, [Fre94]), y en la síntesis de voz se utilizaba un sistema de concatenación de unidades.

Una de las preocupaciones principales en el diseño de DIPLOMAT fue el tratamiento de los errores de las tecnologías de ASR y MT, para producir una aplicación usable con un pequeño entrenamiento por personas que no fueran traductores. Para ello se permitía una pequeña interacción con el usuario, presentando resultados intermedios que permitían corregir posibles errores.

El proyecto TONGUES se basaba en DIPLOMAT, y tenía como objetivo desarrollar un prototipo de sistema de traducción voz a voz que pudiera funcionar en una computadora pequeña. Este fue usado por los US Army Chaplains para comunicarse con refugiados durante abril del 2001 en Zagreb [Fre02].

1.3. TC-STAR

El proyecto TC-STAR fue financiado por la Comisión Europea en el contexto del Sexto Programa Marco. Su objetivo era realizar un esfuerzo durante varios años para lograr avances en las tecnologías de la traducción voz a voz: ASR, MT y TTS. Los integrantes del proyecto eran ITC, RWTH-AACHEN, CNRS-LIMSI, UPC, UKA, IBM, SIEMENS, SRIT, NOKIA, SONY, ELDA y KUN-SPEX.

El proyecto se enfocaba en la traducción voz a voz de habla conversacional sin restricción de dominio de discursos parlamentarios y difusión de noticias en tres idiomas: inglés europeo, español europeo y chino mandarín. Esta tesis se ha desarrollado en el marco del proyecto TC-STAR, en el subproyecto dedicado a la generación de voz en inglés y español europeo.

1.3.1. Resultados obtenidos en ASR

Entre los objetivos a largo plazo del proyecto se encontraba el reconocimiento de voz robusto para diversos estilos de habla, condiciones de grabación y comunidades de usuarios. El sistema debía ser capaz de adaptarse de manera transparente a condiciones particulares.

En el transcurso de los años del proyecto hubo un progreso constante para cada uno de los idiomas, siendo el avance en la reducción de la tasa de palabras erróneas (WER: Word Error Rate) entre los años 2005 y 2006 de un 40 % relativo. La mejor tasa de reconocimiento medida en WER en el año 2007 resultó del 6.9 % para el inglés, 7.4 % para el español y 7.5 % para el chino mandarín. Sin embargo, se concluyó que dicha tasa de reconocimiento debería ser mejorada aún más en el futuro debido a que los sistemas de traducción automática necesitan textos con menos errores para lograr mejores resultados.

Una de los principales causas de los problemas de reconocimiento observadas en los resultados de la última evaluación fue el desequilibrio en el número de locutores masculinos y femeninos. Esto ocurre a causa de que hay más personas de sexo masculino en el Parlamento Europeo.

Finalmente, el análisis de la prestación de reconocimiento de los sistemas del proyecto

TC-STAR discriminando por acento y estilo del habla, indicó que los peores resultados se obtuvieron para hablantes no-nativos o aquellos con un acento muy fuerte. Por ejemplo, la peor prestación se obtuvo para un hablante con un fuerte acento irlandés y una alta velocidad del habla: 19.2 %. La segunda peor prestación correspondió a un hablante húngaro del Parlamento Europeo: 17.7 %.

Además del acento, la fluidez del habla constituyó otro factor con un importante impacto en la tasa de reconocimiento. En algunos casos los hablantes introducen una gran cantidad de disfluencias que hace más dificultosa la tarea de reconocimiento.

1.3.2. Resultados obtenidos en MT

Otro de los grandes objetivos del proyecto TC-STAR fue la traducción efectiva de habla conversacional sin restricciones para grandes dominios de discurso, y la integración efectiva del reconocimiento de voz y la traducción automática en un marco único estadístico.

Uno de los desafíos mayores fue la extensión de los modelos actuales de traducción automática estadística para considerar las múltiples hipótesis de salida producidas por el sistema de reconocimiento automático de voz.

Para el estudio de la traducción automática español-inglés e inglés-español, se hizo uso de tres tipos diferentes de datos como entrada: ASR ROVER, VERBATIM y FTE.

El primero de ellos, denominado ASR ROVER, era la combinación de las salidas de diferentes sistemas de reconocimiento de voz, lo cual proporcionaba una transcripción con el mínimo WER. La salida poseía mayúsculas/minúsculas y puntuación.

La transcripción VERBATIM era realizada en forma manual y fue proporcionada por ELDA (Evaluations and Language resources Distribution Agency). Este tipo de transcripción posee distintos fenómenos de habla espontánea, como es el caso de correcciones, falsos comienzos, etc. En este caso también se proporcionaban mayúsculas/minúsculas y puntuación. Este tipo de transcripción modela la salida de un ASR sin errores.

Finalmente, el último tipo de datos de entrada se denominaba FTE (Final Text Editions), proporcionadas tanto para el Parlamento Europeo como para el Parlamento Español. En este caso muchas oraciones fueron creadas por los servicios de edición del Parlamento, incluyendo puntuación, mayúsculas/minúsculas y la eliminación de las disfluencias introducidas por el habla espontánea.

El análisis de los resultados demuestra que la calidad del texto de entrada es muy importante para una mejor traducción automática. Los sistemas de traducción poseen una mejor calidad de salida utilizando FTE, seguido por el tipo de entrada VERBATIM. Las disfluencias y frases gramaticalmente incorrectas del habla espontánea contribuyen a degradar la prestación de los sistemas de traducción.

Por otra parte, existe una correlación entre la cantidad de errores introducidos por el ASR (WER) y los errores en la traducción. Sin embargo, esta tendencia no es tan marcada como era esperable, siendo necesarios más estudios al respecto en el futuro para analizar su influencia.

1.3.3. Objetivos en TTS

Debido a que uno de los principales objetivos del proyecto TC-STAR era la generación de habla inteligible, expresiva, y que respetara las características del discurso que identificaban al hablante que estaba siendo traducido, era necesario el desarrollo de nuevos modelos para la prosodia, emociones y habla expresiva en general.

Este objetivo motivó la realización de esta tesis para proporcionar nuevos modelos prosódicos para la entonación, duración de fonemas y junturas terminales. En este sentido se detallan en el Capítulo 3 algoritmos para la generación de tales parámetros prosódicos, mientras que en el Capítulo 5 se extienden dichos algoritmos con el uso de la información provista por la señal acústica del hablante del idioma origen para la mejora de la expresividad.

1.4. Objetivos de la tesis

Los objetivos de la tesis son el desarrollo de nuevos algoritmos para el entrenamiento de modelos de generación de prosodia para la conversión texto a voz, y su aplicación en el marco de la traducción voz a voz. Para ello se investigará la posibilidad de mejorar la naturalidad y expresividad de la conversión texto a voz utilizando la prosodia del hablante fuente disponible en el proceso de traducción voz a voz como información adicional.

Con este enfoque se pretende que la voz sintetizada posea diversas características del discurso del hablante fuente. De esta manera, y en combinación con técnicas de conversión de voz, la salida de la conversión texto a voz tendrá tanto la identidad acústica como estilística del hablante fuente.

Las características prosódicas más relevantes son la frecuencia fundamental, la duración de los fonemas, y la posición de las junturas terminales y pausas. Para su generación en la síntesis de voz se utilizarán modelos derivados de métodos de aprendizaje automático en base a datos. Para ello se analizarán los métodos actuales de modelado prosódico para TTS, seleccionando aquellos que puedan generalizarse para el entorno de la traducción y resulten aplicables a varias lenguas. Además, se espera introducir mejoras a los mismos para lograr una mejor calidad y naturalidad.

En los estudios se hará uso de algunas suposiciones para mejorar las condiciones de análisis del funcionamiento de los algoritmos y la medición de su rendimiento. Teniendo en cuenta las limitaciones que presentan los sistemas de reconocimiento automático del habla, traducción automática y conversión texto a voz en lo referente a los errores que introducen y que repercuten en las tareas siguientes, hemos asumido que tanto el ASR como la MT son perfectos. Esta simplificación nos permite analizar mejor el efecto de los modelos propuestos de forma aislada, sin considerar otros aspectos. El estudio de la robustez del sistema frente a errores de ASR/MT requerirá un trabajo posterior fuera del alcance de esta tesis.

1.5. Estructura de la tesis

La tesis está organizada en varios capítulos. En este capítulo hemos realizado una introducción general a la traducción voz a voz y sus componentes, el desarrollo del estado de la cuestión de la traducción voz a voz, y las propuestas en esta tesis para mejorar la calidad y expresividad de la conversión texto a voz en el marco de la traducción voz a voz.

En el Capítulo 2 se describe el estado de la cuestión en las áreas del modelado de la prosodia, tanto en lo que se refiere a la entonación, duración y junturas terminales.

Los nuevos enfoques de entrenamiento para mejorar los resultados obtenidos con técnicas propuestas en la literatura se detallan en el Capítulo 3. Estos enfoques propuestos están aplicados a la predicción de la entonación, la duración segmental y las junturas terminales. Allí se propone una técnica de entrenamiento que combina en un bucle la extracción de los parámetros y la generación del modelo, evitando ciertas suposiciones existentes en la literatura que tienden a degradar el rendimiento de los modelos. Esta técnica se aplica tanto al modelado de la entonación (Sección 3.1.2) como al modelado de la duración segmental (Sección 3.2.3). En lo referente a las junturas terminales, se exploran tres enfoques distintos para el modelado de las mismas, utilizando como unidad de análisis las palabras y los grupos acentuales. Este estudio comparativo permite ver las fortalezas y debilidades de cada una en las mismas condiciones experimentales (Sección 3.3).

La validación experimental de las propuestas se encuentra en el Capítulo 4. Allí se detallan los resultados experimentales usando tanto medidas objetivas (RMSE, correlación, F-measure) como subjetivas (escala MOS de naturalidad y calidad, tasa de error subjetiva).

En el Capítulo 5 se aborda el modelado de la entonación, duración y junturas terminales en el marco de la traducción voz a voz. Teniendo en cuenta que existen más fuentes de información (parámetros acústicos extraídos del hablante origen usando la segmentación provista por el sistema de reconocimiento automático del habla e información de alineamiento y traducción proporcionados por la traducción automática) se propone su uso en conjunto para mejorar la calidad de la prosodia tanto en naturalidad, expresividad y adecuación. La Sección 5.3 describe un algoritmo para la extracción automática de patrones entonativos relacionados entre los idiomas origen y destino. Mediante una técnica de agrupamiento se obtiene de manera automática y no supervisada un conjunto de movimientos tonales que tienen una relación entre los idiomas. En la Sección 5.4 se estudia la sincronización del audio traducido con el video del hablante origen, que indirectamente implica una transferencia del ritmo. Se propone un conjunto de técnicas de sincronización para coordinar el mensaje con los movimientos del orador. Finalmente, la Sección 5.5 describe la transferencia de junturas terminales del hablante origen en el idioma destino. A través de una técnica de transferencia de pausas del idioma origen al destino es posible preservarlas con el objeto de conservar el significado del mensaje.

En el Capítulo 6 se detallan las conclusiones de este trabajo sobre las aportaciones realizadas en el modelado prosódico y en la transferencia de la prosodia en la traducción oral. Allí también se proponen direcciones futuras para continuar el progreso en el área.

La arquitectura de Ogmios, el conversor texto a voz utilizado en el desarrollo de esta tesis, se describe en el Apéndice A. Allí se detallan sus módulos de **Análisis del texto**,

Generación de la prosodia, y **Generación de la voz**. En este apéndice también se describe en forma resumida el proceso de generación de una voz sintética en base a un conjunto de grabaciones.

En el Apéndice B se explican un conjunto de herramientas estadísticas utilizadas en la evaluación de los diferentes algoritmos propuestos en esta tesis: error cuadrático medio, coeficiente de correlación Pearson, box-plots, y el test de Wilcoxon.

Tanto el corpus monolingüe utilizado para la generación de la voz en TTS, como el corpus bilingüe para el estudio de la transferencia de la prosodia del idioma fuente al idioma destino, se encuentran descritos en el Apéndice C. Allí se detallan tanto las condiciones de grabación como el contenido de los diferentes corpus.

Finalmente, los resultados de la investigación durante el desarrollo de esta tesis se vieron reflejados en publicaciones en diversas conferencias y en el aporte al proyecto TC-STAR, tal como se detalla en el apéndice D sobre publicaciones.

Capítulo 2

Modelado prosódico en los sistemas de síntesis de voz

Según el Diccionario de la Real Academia Española disponible en internet (http://www.rae.es) una de las acepciones de la palabra prosodia es la siguiente:

prosodia *(Del lat. prosodia, y este del gr. προσωδια).*

Parte de la fonología dedicada al estudio de los rasgos fónicos que afectan a unidades inferiores al fonema, como las moras, o superiores a él, como las sílabas u otras secuencias de la palabra u oración.

A través de diferentes recursos prosódicos, tales como la entonación, ritmo, intensidad y pausas, se estructura el habla y el discurso, y esto constituye uno de los usos más importantes de la prosodia. En general, se puede afirmar que no es posible entender una oración sin el uso de dichos recursos debido a la gran información que proporcionan. Además, cuando hablamos no transmitimos solamente el mensaje contenido en las palabras, sino también importante información acerca de nuestra identidad (género, dialecto, edad, origen social) y nuestro estado de ánimo, emociones e intención.

La prosodia del hablante también está condicionada por el área geográfica donde nació o bien donde vive. Es muy común observar estas "huellas digitales". Por ejemplo, los italianos hablan como cantando, los bolivianos usan un ritmo lento o los ciudadanos de la provincia de Córdoba (Argentina) ponen un doble acento en algunas palabras. En algunos casos estas características son consecuencia de los orígenes de las migraciones. Muchos dicen que el argentino es un italiano hablando en español, o incluso, un importante escritor argentino fue más allá con sus afirmaciones:

> *El gran escritor Jorge Luis Borges ha evocado que: el argentino es un italiano que habla español, se comporta como un francés pero quisiera ser inglés.*

En muchas situaciones el estado mental del hablante también se ve reflejado en la

prosodia y no solamente en las palabras del mensaje. En algunos casos las personas no usan palabras para explicar como se sienten, sino que usan elementos prosódicos para indicarlo. Una disminución en el rango de la frecuencia fundamental o un ritmo más lento es un indicador de tristeza o depresión. Por otra parte, una subida en el rango del tono y un ritmo más rápido puede implicar felicidad.

Es importante remarcar que las personas no hablan solamente para transmitir información al locutor, sino también para causar un efecto en él. En algunas ocasiones los recursos prosódicos se usan para poner énfasis en aspectos importantes del mensaje, forzar a una persona a obedecer una orden mediante gritos o hacer sentirla más calma con un ritmo y una intensidad del habla más apacible.

Esta enumeración de los usos de la prosodia indica el grado de importancia de la misma. Su generación en forma adecuada en la conversión texto a voz es necesaria para lograr naturalidad y expresividad. En la siguiente sección se describe la estructura básica de un sistema de conversión de texto en habla, con el objeto de comprender todas las tareas involucradas en el proceso de convertir un texto en voz. De esta manera será posible visualizar los elementos que impactarán en el modelado y la generación de la prosodia.

Como hemos visto en la introducción un sistema de conversión texto a habla se compone de tres módulos: procesamiento del texto, generación de la prosodia y generación de la voz sintética. Esta tesis se centrará en el módulo de generación prosódica. Sin embargo, para poder situar a la prosodia en el contexto de la síntesis de voz, en los siguientes apartados (secciones 2.1.1, 2.1.2 y 2.1.3) se dará una breve explicación de los diferentes módulos de un conversor texto a voz y los enfoques utilizados en la literatura para su implementación.

A continuación en los apartados 2.2, 2.3 y 2.4 se hará una introducción a cada uno de los elementos de la prosodia estudiados en esta tesis: entonación, duración y junturas terminales, respectivamente. En cada una de estas secciones se detallarán algunos de los modelos propuestos en la literatura para su generación en el contexto de la conversión texto a voz.

2.1. Conversión de texto en habla

El proceso de conversión texto a voz se divide en tres módulos básicos: procesamiento del texto, generación de la prosodia y generación de la voz sintética. El objetivo de esta sección es describir brevemente el funcionamiento de un sistema de conversión texto a voz y ver la importancia de la prosodia en el sistema, ya que este es el foco de la tesis.

2.1.1. Procesamiento del texto

El procesamiento de texto es una tarea muy compleja que incluye un conjunto de problemas que son altamente dependientes del idioma y del dominio, tales como la normalización del texto y la transcripción fonética. Por ello son necesarios enfoques particulares para cada idioma con el objeto de convertir un texto con un formato particular en una se-

cuencia de unidades segmentales (por ejemplo: fonemas), suprasegmentales (por ejemplo: sílabas), palabras, oraciones y párrafos, aptos para su utilización por parte de los módulos de generación de la prosodia y de la voz.

Normalización del texto

Una de las primeras tareas en el procesamiento del texto es la normalización, que convierte la totalidad del texto a una forma textual convencional. Los dígitos y números se deben expandir en palabras teniendo en cuenta la información que transportan. Por ejemplo, *128* se debe expandir como *ciento veintiocho*. Sin embargo, *1816* es un caso más complejo, porque se puede expandir en forma diferente si representa un año en un texto en inglés: *eigthteen sixteen*. Aparecen más problemas con el símbolo "/": *1/2* se puede expandir como *un medio* si es una fracción o como *uno de enero* si es una fecha. Hay también otros símbolos que combinados con números convierten a la expansión en una tarea muy difícil, por ejemplo, los números telefónicos: (+54 223) 483-3893.

Las abreviaturas también deben recibir un tratamiento especial. La principal dificultad aparece porque algunas abreviaturas se pronuncian tal como están escritas (NATO, RAM) mientras que otras se pronuncian letra por letra (MGM, PP), o bien una combinación de ambas (PSOE, MPEG). En algunos casos es posible predecir la pronunciación de una abreviatura, pero en muchos otros casos es el resultado de un consenso público imposible de predecir. En tales situaciones es necesario tratar a las abreviaturas como casos especiales.

Existe un cúmulo de situaciones que también se deben tener en cuenta en función del origen del texto:

- *Formato del texto.* El texto de entrada puede contener muchos detalles de formato, tales como títulos, información de secciones, pie de página, referencias, etc. Es también posible que el texto sea parte de un periódico o una revista, y por lo tanto contendrá columnas, cajas, tablas, avisos, etc. El tratamiento de alguno de estos aspectos no son específicos de los sistemas de TTS. Ellos pueden depender de etapas previas, tales como software de reconocimiento óptico de caracteres.

- *Hipertexto.* El texto en internet se encuentra en formato de hipertexto. Las etiquetas HTML y los enlaces deben ser tratados cuidadosamente con el fin de obtener un texto correcto. Los detalles de diseño pueden ocasionar serias dificultades para obtener un texto correcto si este se encuentra ubicado espacialmente en una forma complicada para la capacidad de procesamiento de un ordenador. En HTML existen algunas etiquetas tales como $< p > ... < /p >$ que son útiles para delimitar párrafos y ayudar tanto al visualizador de páginas como al sistema de TTS. Además, etiquetas adicionales se pueden usar por parte del sintetizador para mejorar la calidad de la voz generada. Por ejemplo, se pueden utilizar etiquetas $< happy >...< /happy >$ para delimitar un texto que debe ser expresado con felicidad, o bien hacerlo para una pregunta $< quest >...< /quest >$. Las características de la voz y su idioma también se pueden controlar de la misma manera con etiquetas $< gender = female >$ o $< lang = spanish >$. Algunas palabras y nombres comunes tienen pronunciaciones particulares que se pueden corregir con el mismo tipo de etiquetas. Etiquetas de

énfasis pueden usarse para dar un acento particular a alguna palabra en la oración [Spr98b, Bur04, Hun00, EML].

- *E-mail, SMS y chat.* Otra aplicación de la tecnología de voz es la lectura de e-mails, SMS y chats. El amplio uso de estos canales de comunicación ha introducido nuevos códigos sociales de comunicación. Los "emoticons" son uno de estos códigos aceptados mundialmente. Cada día es más común encontrar símbolos tales como :) y :D en los mensajes. Tales símbolos deben recibir una interpretación adecuada por parte del sintetizador. Estados tales como felicidad, tristeza, aburrimiento, o bien fillers tales como risas, sonrisas, gruñidos, etc. Por otra parte, la limitación en el número de caracteres de los SMS's (y el deseo de evitar la división del mensaje y el subsecuente incremento en el costo del mismo) conduce al uso de abreviaciones en ciertas palabras: be se convierte en b, see en c, are en r, you en u, for en 4, etc. Como consecuencia, un mensaje tan simple como *"Sorry I forgot to phone you.I will see you tomorrow"* se puede convertir en el críptico mensaje *"soz i 4gt 2 fon u.i c u 2moz"* [Bea10]. [1]

Transcripción fonética

Es bien conocido que la pronunciación de las palabras difiere de la forma en que se encuentran escritas. Como una consecuencia de ello, el principio de correspondencia "un caracter"→ "un fonema" frecuentemente no es aplicable. Para darse cuenta de ello uno debe observar que [SVNY97]:

- Un simple caracter puede corresponder a dos fonemas, como es el caso de la x /ks/, o bien a ninguno, como ocurre con la e de *surely*.
- Muchos caracteres pueden corresponder a un solo fonema, como es el caso de la ch /ʃ/, o incluso a ninguno, como gh en la palabra inglesa *though*.
- Un caracter, o una secuencia de ellos, se pueden pronunciar de diferentes maneras en función de los contextos derecho y/o izquierdo, tal como ocurre con la c: /s/ en anciano y /k/ en anclaje.
- Por otra parte, un fonema puede aparecer para caracteres diferentes, como sucede con /ʃ/: sh en *dish*, t en *action*, y c en *ancient*.

En general, la pronunciación de palabras aisladas se puede encontrar en un diccionario de transcripción fonética. Para ciertos verbos conjugados, o sustantivos femeninos y/o plurales, su transcripción se puede deducir mediante un conjunto de reglas. En el caso de existir excepciones, estas se deben explicitar mediante un conjunto de reglas específicas o bien a través de un diccionario auxiliar.

En un conversor texto a voz la tarea de convertir un texto en los fonemas constituyentes se denomina conversión de grafemas en fonemas, y puede incluir también la colocación

[1] Nota: Otro uso común es el reemplazo de *orr* con la abreviatura *oz*. De esta manera, *sorry* se convierte en *soz* y *tomorrow* se transforma en *tomoz*. Este último incluso puede abreviarse como *2moz*.

del acento léxico y la separación en sílabas. Esta tarea es dependiente del idioma, y su dificultad puede variar, tal como se observa si se comparan el Español y el Inglés.

El idioma Español posee reglas claras para realizar la transcripción fonética de una palabra. Por ejemplo, en el artículo de Moreno et al. [Mor98] se explica que es posible una transcripción canónica del español, correspondiente a la variante central de España (Castellano), usando un conjunto de reglas. La alteración de algunas de dichas reglas permite la transcripción fonética de otros dialectos, tanto de la península ibérica, baleares, y países latinoamericanos.

En cambio, en el idioma Inglés no se puede realizar una transcripción fonética basada en reglas debido a que no es posible encontrar un número finito de ellas. En consecuencia, para todos aquellos idiomas que no tienen la posibilidad de transcripción fonética por reglas, se utilizan otros enfoques basados en diccionarios y reglas aprendidas en forma automática.

Los diccionarios son listados de palabras donde se encuentra la transcripción fonética de cada una de ellas. Es posible que algunas palabras posean varias transcripciones fonéticas posibles debido a motivos no dialectales. Por ejemplo, en algunas situaciones la pronunciación depende del significado de la palabra (por ejemplo: dessert). Debido a esto, es esencial el uso de información semántica y morfosintáctica para lograr una correcta pronunciación.

La pronunciación de una cierta palabra también puede cambiar debido al contexto. Esto es fácil de ver cuando se comparan frases al final y al comienzo, y por ello la pronunciación de *the* depende del fonema inicial de la siguiente palabra. Las palabras compuestas también son problemáticas. Por ejemplo, los caracteres “th” en “mother” y en “hothouse” se pronuncian de manera diferente. Algunos sonidos también pueden ser sonoros o sordos en contextos diferentes. Por ejemplo, el fonema /s/ es sonoro en la palabra dogs, mientras que es sordo en la palabra cats [All87, Lem99].

Aquellas palabras que necesitan una transcripción fonética y no se encuentran en el diccionario, reciben un tratamiento diferenciado como palabras fuera de vocabulario. Existen en la literatura una gran variedad de propuestas para el aprendizaje automático de la transcripción de estas palabras, tales como pronunciación por analogía, pronunciación por reglas, o diversos enfoques probabilísticos [Bis08].

Dentro de las palabras fuera de vocabulario se encuentran generalmente los nombres propios. Encontrar su correcta pronunciación, especialmente cuando provienen de otros idiomas, es comúnmente una de las tareas más difíciles para cualquier sistema de TTS. Alguno de los nombres comunes, tales como Nice y Begin, son ambiguos cuando se encuentran al comienzo de una oración o en los títulos. Por ejemplo, la oración “Nice is a nice place” es muy problemática porque la palabra “Nice” se puede pronunciar como /nis/ o /nais/. Algunos nombres y lugares tienen también una pronunciación especial, tales como Leicester y Arkansas. Para su correcta pronunciación, este tipo de palabras se debe incluir en un diccionario especial de excepciones. Desafortunadamente, está claro que no hay manera de construir una base de datos que contenga todos los nombres propios que pueden llegar a aparecer.

En muchas ocasiones, como parte de la transcripción fonética se genera también una agrupación de los fonemas en sílabas, con el objeto de proporcionar información suprasegmental útil para las tareas de generación de la prosodia y voz. Las sílabas se consideran como una de las unidades más básicas en el habla de muchos idiomas. Los niños aprenden a producir sílabas mucho antes que puedan decir una palabra de su lengua materna. Otra situación donde se observa la importancia de las sílabas es en algunas personas con problemas en el habla específicos. En ellas todavía se podrá observar organización silábica incluso en habla defectuosa.

Una sílaba se define como una unidad del lenguaje hablado más grande que un sonido del habla (fonema), y que está constituido de hasta tres componentes: un núcleo, el cual consiste de una vocal simple o bien una consonante silábica [2], acompañado opcionalmente por una o más consonantes. Las consonantes que preceden al núcleo son llamadas ataque, mientras que aquellas que se encuentran luego del núcleo se denominan coda. El núcleo y la coda a veces se los considera conjuntamente para formar la rima.

Los sistemas de TTS tienen un componente que realiza la división de las palabras en sílabas (silabificación). El procedimiento es dependiente del lenguaje. Por ejemplo, en algunos idiomas tales como el español, las reglas de silabificación son simples. Por otra parte, los idiomas como el inglés tienen reglas complejas que son más difíciles de escribir. En tales situaciones puede utilizarse un diccionario conjuntamente con reglas aprendidas por métodos estadísticos.

2.1.2. Modelado prosódico

Tal como hemos explicado en la introducción de este capítulo, la prosodia involucra la parte de la comunicación humana que expresa las emociones, enfatiza palabras, muestra la actitud del hablante, divide una oración en frases, y también implica el ritmo y la entonación en el habla.

Por ello, en un conversor texto a voz se incluirán módulos para la generación de los diferentes componentes de la prosodia. Cada uno de estos módulos hará un uso particular de la secuencia de unidades segmentales (por ejemplo: fonemas), suprasegmentales (por ejemplo: sílabas), palabras, oraciones y párrafos.

Entre los módulos para la generación de la prosodia se pueden encontrar los encargados de la entonación, la duración, las junturas terminales y las pausas, que serán los tratados en esta tesis. En la literatura también se pueden encontrar otros módulos relacionados a otros aspectos de la prosodia, como es el caso de la *voice quality*.

En este apartado no se incluye una descripción de los módulos de entonación, duración, junturas terminales y pausa, ya que son tratados en el resto de la tesis.

[2] Sonido consonántico que puede desempeñar la función de núcleo silábico. En español no existen consonantes silábicas; sin embargo, en otras lenguas como el inglés, las líquidas (laterales y vibrantes) y las nasales pueden funcionar como núcleos silábicos [Tra05].

2.1.3. Generación de voz artificial

La generación de voz artificial es un campo que ha evolucionado constantemente desde el siglo XVIII con los trabajos de Kratzestein y Wolgang von Kempelen.

Existen en la literatura diferentes enfoques para la generación de voz artificial, que van desde la aplicación de la física al aparato fonatorio, hasta el uso de enfoques estadísticos.

Síntesis articulatoria

Uno de los enfoques más directos para la generación de voz artificial es la síntesis articulatoria, ya que simula el sistema de producción del habla, es decir, el funcionamiento de los órganos del aparato fonatorio humano. De esta manera se puede producir una voz sintética de alta calidad emulando los diferentes articuladores y las cuerdas vocales.

El sintetizador de Wolgang von Kempelen es el más antiguo, y utilizaba un enfoque que puede ser considerado de síntesis articulatoria. Mediante un conjunto de tubos reproducía sonidos que se podían reconocer como habla. Muchos sintetizadores del habla articulatorios modernos utilizan también modelos de tubos acústicos. Un tubo de forma compleja puede ser simulado con un conjunto de tubos uniformes más pequeños, y mediante las propiedades de propagación de sonido de tales sistemas más simples, es posible construir un modelo general más complejo.

Esta técnica presenta dos tipos de dificultades. La primera de ellas es decidir los parámetros de control en base a la especificación. Los parámetros articulatorios no pueden ser deducidos de grabaciones, y por lo tanto son necesarias técnicas intrusivas para hallarlos. En la actualidad se han producido grandes avances en el sensado del movimiento de los articuladores usando tanto EMG (ElectroMioGrafía) como MRI (Magnetic Resonance Imaging). Esto ha contribuido a la producción de mejores modelos articulatorios usando la información de los músculos y modelos en tres dimensiones.

La otra dificultad se encuentra en el momento de decidir el grado de precisión necesaria para producir un modelo que se ajuste a la fisiología humana, pero que además sea manejable desde el punto de vista del diseño y el control. En este punto es necesario destacar que este enfoque es uno de los que requiere la mayor carga computacional debido a la matemática envuelta en el proceso de generación de la voz [Krö92, Rah93]. Por lo tanto, ha sido uno de los métodos más postergados inicialmente debido al limitado poder de cálculo de los ordenadores.

Es tal la dificultad que presentan ambos problemas, que los mejores sistemas de síntesis articulatoria tienen una calidad pobre comparados con los mejores sistemas de síntesis que utilizan otros enfoques. Debido a esto, la síntesis articulatoria ha sido abandonada como técnica para la generación de habla de alta calidad para propósitos ingenieriles.

Síntesis por formantes

Uno de los métodos más usados durante la década de los 80 fue la síntesis por formantes. La misma adopta un enfoque acústico-fonético y modular para la generación de la

voz, donde se usa un modelo de tubos acústicos con elementos de control que pueden ser fácilmente relacionados con propiedades acústico-fonéticas. Para la síntesis basada en formantes son necesarias un conjunto de reglas para determinar los valores de los parámetros utilizados para sintetizar una oración dada [All87].

En un esquema típico de un sintetizador de formantes el sonido se genera usando una fuente periódica para los sonidos sonoros, y ruido blanco para los sonidos sordos. Las cavidades nasal y oral se modelan como sistemas paralelos de filtros. La señal pasa a través de la cavidad oral, pero también puede hacerlo por la cavidad nasal, en caso de que así lo requiera un sonido nasal. Finalmente, estas componentes se combinan y pasan por un filtro de radiación, que simula las características de propagación de los labios y la nariz.

En general se utilizan dos estructuras de filtros, paralela y en cascada, obteniéndose la mejor calidad usando una combinación de ellas. Al menos son necesarios tres formantes para producir habla inteligible, pero en general se usan cinco para obtener voz de alta calidad. Cada formante se modela con un resonador de dos polos, lo cual permite determinar tanto la frecuencia del formante como su ancho de banda [Don96].

En la Figura 2.1 se puede observar un sintetizador por formantes completo: el sintetizador Klatt. Este es uno de los sintetizadores de formantes más sofisticados que se han desarrollado.

Síntesis por concatenación

La concatenación de segmentos pregrabados es probablemente la manera más fácil de producir habla sintética inteligible y natural. Sin embargo, debido a su principio de funcionamiento, los sintetizadores por concatenación están limitados a producir la voz de un hablante en particular, con grandes requisitos de capacidad de memoria.

Uno de los aspectos más importantes en la síntesis por concatenación es la búsqueda de la unidad apropiada a concatenar. La selección es un compromiso entre la utilización de unidades largas y cortas. Las unidades largas son más naturales, debido a que hay menos puntos de unión y un mejor control de la coarticulación. Sin embargo, la cantidad de memoria necesaria es más grande, debido a la cantidad posible de combinaciones en un idioma. Por otra parte, la utilización de unidades cortas produce menos requerimientos de memoria, pero la recolección de unidades apropiadas es más compleja, y la calidad y naturalidad se ven seriamente degradadas. En la actualidad, los sistemas usan una combinación de las diferentes unidades, utilizando las más largas cuando están disponibles. Las unidades más cortas (semifonemas y fonemas) son utilizadas para cubrir algunos casos poco frecuentes.

Uno de los métodos más usados para la concatenación de unidades es PSOLA (Pitch Synchronous Overlap Add). Dicho método fue desarrollado por France Telecom (CNET), y es utilizado por muchos sistemas de síntesis comerciales. Existen varias versiones del algoritmo PSOLA, pero en general todas ellas se basan en el mismo principio. La versión en el dominio del tiempo es TD-PSOLA, y es la más comúnmente usada debido a su eficiencia computacional [Kor97]. El algoritmo básico consiste en tres pasos [Cha89, Val91]: análisis de la señal original dividiéndola en tramos solapados sincronizados con el pitch, modifi-

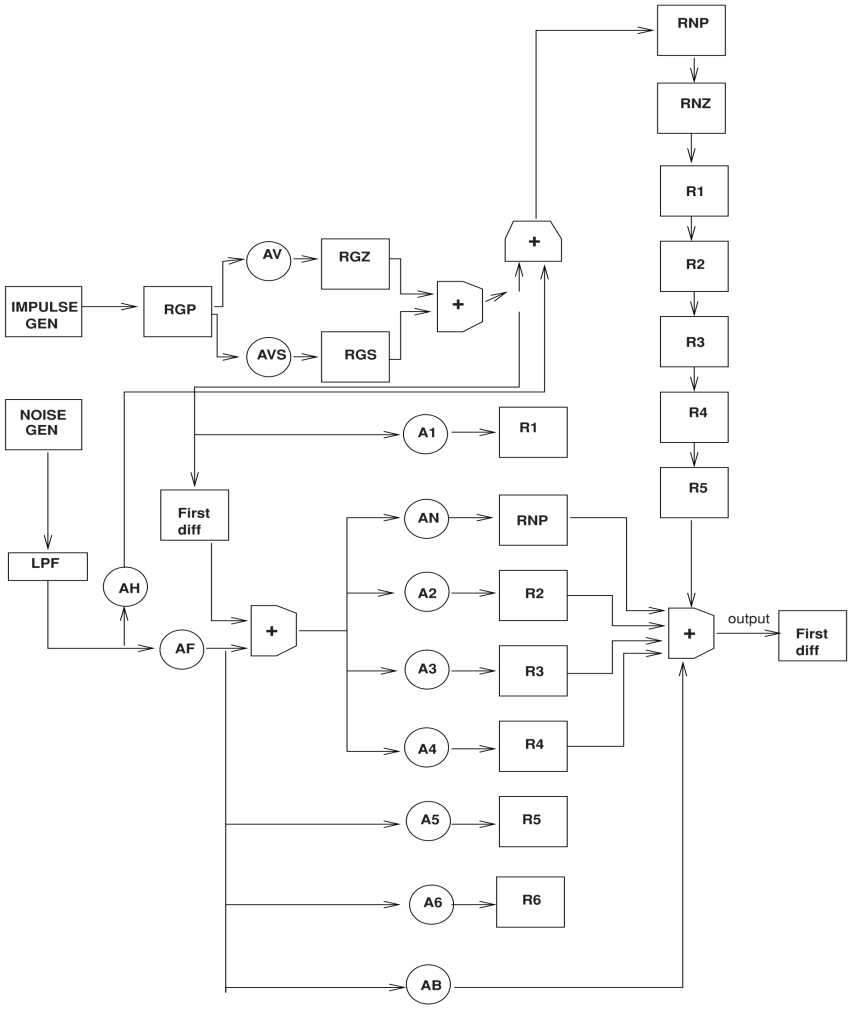


Figura 2.1: Sintetizador Klatt.

cación de la señal analizada, y generación de la señal sintética mediante la recombinación por suma solapada [Mou90].

Síntesis por selección de unidades

Los sistemas de síntesis por concatenación asumen que las variaciones acústicas que se pueden producir en un fonema son atribuíbles a diferencias en el tono y la duración. Además también consideran que los algoritmos de procesamiento de señal son capaces de realizar todos los cambios necesarios en el tono y la duración sin incurrir en una pérdida de naturalidad [Tay09].

Estas afirmaciones se convierten en los factores limitantes de la calidad de la síntesis que se obtiene en la práctica con tales sistemas. A pesar que existen un gran número de trabajos para desarrollar algoritmos de procesamiento se señales más eficientes, no resulta

suficiente introducir cambios en el tono y la duración para lograr una voz natural, ya que otros factores tales como la energía, las características dinámicas de la articulación y la *voice quality* también ejercen una influencia importante.

La observación de estas debilidades llevaron al desarrollo de un conjunto de técnicas conocidas como selección de unidades. Estas técnicas usan segmentos de voz con una variedad más rica de características acústicas, con el objeto de capturar las distintas variantes que pueden existir y depender menos de la modificación de la señal.

El objeto de este tipo de técnicas es utilizar un conjunto de unidades que pueden ser agrupadas usando criterios lingüísticos, y que presentan variaciones en lo relativo a la prosodia y otras características. Durante el proceso de síntesis, un algoritmo selecciona una unidad entre todas las disponibles, con el objeto de encontrar la mejor secuencia de unidades que se ajuste a las especificaciones, minimizando efectos no deseados tales como discontinuidades.

Entre los trabajos relacionados con este enfoque se encuentra el de Hunt y Taylor [Hun96], en donde los autores explican el uso de dos funciones de costo: objetivo y concatenación. El costo objetivo es una estimación de la diferencia entre la unidad presente en la base de datos y la unidad objetivo que deberá representar. El costo de concatenación es una estimación de la calidad de la unión de dos unidades seleccionadas de la base de datos. En consecuencia, el algoritmo propuesto por estos autores tienen en cuenta tanto el ajuste a las especificaciones como minimizar discontinuidades que podrían surgir en la concatenación.

Otros artículos que avanzan en una síntesis más expresiva son los de Pitrelli [Pit06] y Steiner [Ste10], donde se explican en forma detallada el funcionamiento de sistemas de conversión texto a voz de habla expresiva usando selección de unidades.

Síntesis por HMM

En los últimos años ha habido un creciente interés en la generación de voz sintética usando Modelos Ocultos de Markov (HMM: Hidden Markov Models), debido a que han sido ampliamente estudiados y es posible obtener voz sintética natural. En este método, tanto el espectro, la frecuencia fundamental y la duración segmental son modelados usando un marco HMM unificado.

Una ventaja resultante de este enfoque de modelado basado en parámetros es la flexibilidad con respecto a los métodos basados en selección de unidades y concatenación. En síntesis por HMM es posible usar técnicas de adaptación e interpolación de modelos para controlar los parámetros y las características asociadas al habla [Yam07, Nos07]. Además, Yamagishi et al. [Yam08] demostraron que el modelado usando HMM también es robusto a condiciones de grabación que son perjudiciales para otros enfoques, tales como la concatenación de unidades.

En la síntesis de voz los HMM permiten modelar tanto los parámetros de excitación como los espectrales usando HMM dependientes del contexto [Yos99]. En el proceso de síntesis, los parámetros espectrales y de excitación son generados a través de los HMM. Luego, la excitación es filtrada usando los parámetros espectrales para generar la voz

sintética [Tok95].

Una de las principales limitaciones del sistema básico es que el habla posee un zumbido debido a que la técnica utilizada produce sonido con un estilo similar al obtenido por un vocoder. Para minimizar este problema se han propuesto un conjunto de técnicas, tales como STRAIGHT [Kaw99]. Además, para mejorar el modelado de la duración se ha propuesto un modelado acústico basado en modelos semi-Markov [Zen04], y para reducir la monotonía de la voz se generan los parámetros incluyendo en la función de coste de la frase generada la varianza global de la parámetros [Tod05].

Síntesis por predicción lineal

La síntesis por predicción lineal fue un método diseñado originalmente para sistemas de codificación del habla. Sin embargo también se utiliza en sistemas de síntesis de voz por concatenación y en la síntesis estadística por su utilidad para realizar manipulaciones de la frecuencia fundamental y la duración de los fonemas. Esta técnica se basa en los mismos principios que la síntesis por formantes, donde una señal de excitación se pasa por un filtro para obtener la voz sintética.

En este caso, el filtro solamente está constituido de polos, y se modela como una secuencia de coeficientes que minimizan el error de predicción lineal de la señal:

$$e(n) = y(n) - \sum_{k=1}^{p} a(k)y(n-k) = y(n) - \hat{y}(n) \tag{2.1}$$

El principio básico de la predicción lineal se basa en el hecho que la muestra $y(n)$ puede ser predicha usando un conjunto p de muestras $y(n-1)$ a $y(n-p)$ a través de una combinación lineal, y que presentará un error $e(n)$ llamado señal de residuo.

En la fase de síntesis la señal de excitación se puede aproximar por un tren de impulsos para los sonidos sonoros y por ruido aleatorio para sonidos sordos. Dicha señal de excitación es amplificada y filtrada por el filtro digital cuyos coeficientes son $a(k)$.

La principal deficiencia del algoritmo original de predicción lineal es la representación del tracto vocal como un modelo que tiene solamente polos, lo cual es una modelización pobre para aquellos sonidos que poseen antiformantes, como es el caso de las consonantes nasales y las vocales nasalizadas.

Otro aspecto a tener en cuenta es que el modelado de la señal en base a un conjunto de predictores lineales implica que el filtro todos-polos también modelará el filtro glotal. En consecuencia, el filtro modela tanto el tracto vocal como el filtro glotal. Esto es una gran diferencia con respecto al sintetizador basado en formantes que posee un filtro glotal para producir en una forma más precisa y realista la señal glotal de volúmen-velocidad. La simplicidad de la señal de excitación en la síntesis por predicción lineal (mediante impulsos) produce un sonido resultante metálico, semejante a un zumbido, que degrada la calidad y la naturalidad de la voz resultante.

La calidad de la síntesis por predicción lineal se considera en general pobre. Sin embargo, algunas modificaciones y extensiones del modelo básico mejoran la calidad obtenida,

tales como WLP (Warped Linear Prediction) [Lai94, Kar98], Multipulse Linear Prediction (MLPC) [Ata82], Residual Excited Linear Prediction (RELP) [Mag74], Code Excited Linear Prediction (CELP) [Sch85] y Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) [Kaw99].

Síntesis por sinusoides

Los modelos sinusoidales también son utilizados en sistemas de síntesis de voz por concatenación y en la síntesis estadística por su utilidad para realizar manipulaciones de la frecuencia fundamental y la duración de los fonemas. Estos modelos se basan en el principio de que la señal del habla se puede representar como una suma de sinusoidales con amplitudes y frecuencias variantes en el tiempo [McA86, Mac96, Kle98]. En el modelo sinusoidal básico, la señal del habla $s(n)$ se modela como la suma de un pequeño numero L de sinusoides

$$s(n) = \sum_{i=1}^{L} A_i(n) cos(\omega_i(n)n + \phi_i) \tag{2.2}$$

donde $A_i(n)$ y $\phi_i(n)$ representan la amplitud y fase de cada componente sinusoidal con la frecuencia $\omega_i(n)$. Para encontrar estos parámetros se utilizan estimaciones de los picos espectrales en una ventana de la señal. Esta estimación periódica es adecuada para sonidos sonoros, tales como vocales y consonantes sonoras. Sin embargo, la representación de sonidos sordos es más problemática.

Los modelos sinusoidales son utilizados frecuentemente en la síntesis de voz de canto [Mac96, Mac97]. Este tipo de síntesis difiere de la conversión texto a voz en que la entonación y las cualidades musicales son más relevantes que la inteligibilidad del mensaje.

El modelo armónico más ruido (HNM: Harmonic/Noise Model) propuesto por Stylianou constituye una mejora a la síntesis por sinusoides [Sty01]. Tal como el nombre lo indica, este modelo esta compuesto por dos componentes, una armónica y una de ruido. La componente de ruido es más sofisticada que en los modelos mencionados anteriormente, ya que considera el hecho que puede haber patrones temporales específicos en el habla real. Por ejemplo, las plosivas tienen una componente de ruido que varía a lo largo del tiempo, y sería incorrecto forzar una uniformidad, ya que se perderían importantes detalles de la señal real. Dicha componente se modela con la siguiente expresión: $s(t) = e(t)\,(h(t,\tau) \otimes b(t))$, donde $b(t)$ es ruido blanco gaussiano, $h(t,\tau)$ es un filtro espectral aplicado al ruido, y $e(t)$ es una función que proporciona el patrón temporal correcto.

2.1.4. Importancia de la prosodia en la generación de voz

La generación de la prosodia es extremadamente importante en la generación de voz sintética, ya que ejerce una gran influencia en los distintos aspectos de la misma.

En el caso de la generación de voz por métodos que separan el tracto vocal de la señal de excitación (síntesis articulatoria, síntesis por formantes y síntesis por predicción lineal)

la melodía de la prosodia ejerce una fuerte influencia en la frecuencia de la excitación de los sonidos sonoros, mientras que la intensidad afecta la amplitud de los formantes y la ubicación de las frecuencias de resonancia con respecto a la frecuencia de la excitación. Finalmente, el ritmo determinará las trayectorias de los frecuencias de resonancia del tracto vocal y su velocidad de variación.

Por otra parte, la prosodia ejerce una influencia a diferentes niveles en la síntesis por concatenación. En el caso de la selección de unidades, la prosodia se utiliza para encontrar una secuencia óptima de unidades en la base de datos que minimicen la necesidad de hacer cambios prosódicos, tales como la modificación de la duración o la melodía. Además, en algunos casos la prosodia también es utilizada para realizar una modificación en la duración de los fonemas o la melodía. En general, estos cambios no son deseados y se prefiere evitarlos aumentando la cobertura prosódica de la base de datos.

En la síntesis por HMM los modelos ocultos de Markov permiten modelar la prosodia intrínsecamente. En un HMM estándar, las probabilidades de transición determinan las características de duración del modelo, y las duraciones generadas tienen una densidad de probabilidad exponencial. Mediante experimentos se puede determinar que las duraciones de los fonemas tienen una densidad de probabilidad gaussiana en la escala logarítmica, y es por ello que el modelado de duraciones exponencial de los HMM es inexacto. Por ello se han propuesto en la literatura *Hidden Semi-Markov Models* (HSMM) para reemplazar las probabilidades de transición por un modelo de duración explícito gaussiano [Lev86, Yam04, Zen05].

2.2. Entonación

Uno de los componentes más importantes de la prosodia es la entonación, o la melodía de una oración, la cual lleva información acerca del hablante y el mensaje. Todos los hablantes de una lengua conocen el grupo de contornos de entonación que se usan para expresar una variedad de significados. Incluso, los hablantes son capaces de distinguir aquellos contornos que pertenecen a su lengua de aquellos que no.

Físicamente, la entonación se produce por variaciones de la frecuencia fundamental (F0) del habla, que es la frecuencia de los pulsos glotales generados por la vibración de las cuerdas vocales en los segmentos sonoros.

El tono del habla es el correlato perceptual de la f_0. Las escalas psicoacústicas de tono son lineales solamente para frecuencias relativamente bajas. Sin embargo, se asume que hay una correlación lineal entre el tono y la F0 en los rangos de frecuencia que son relevantes para el habla sonora de hombres y mujeres (50Hz-250Hz y 120Hz-400Hz, respectivamente).

La entonación generalmente se representa usando un gráfico de dos dimensiones donde el eje de las abscisas es el tiempo y el eje de las ordenadas es la frecuencia fundamental.

En la Figura 2.2 se puede observar la entonación de la oración "*¿Cómo se llamaba el caballo de Calígula?*". En el contorno se ven en ocasiones picos, de los cuales algunos corresponden a sílabas acentuadas. También es posible observar la tendencia a la declinación presente en muchos idiomas, y la falta de información sobre la frecuencia fundamental

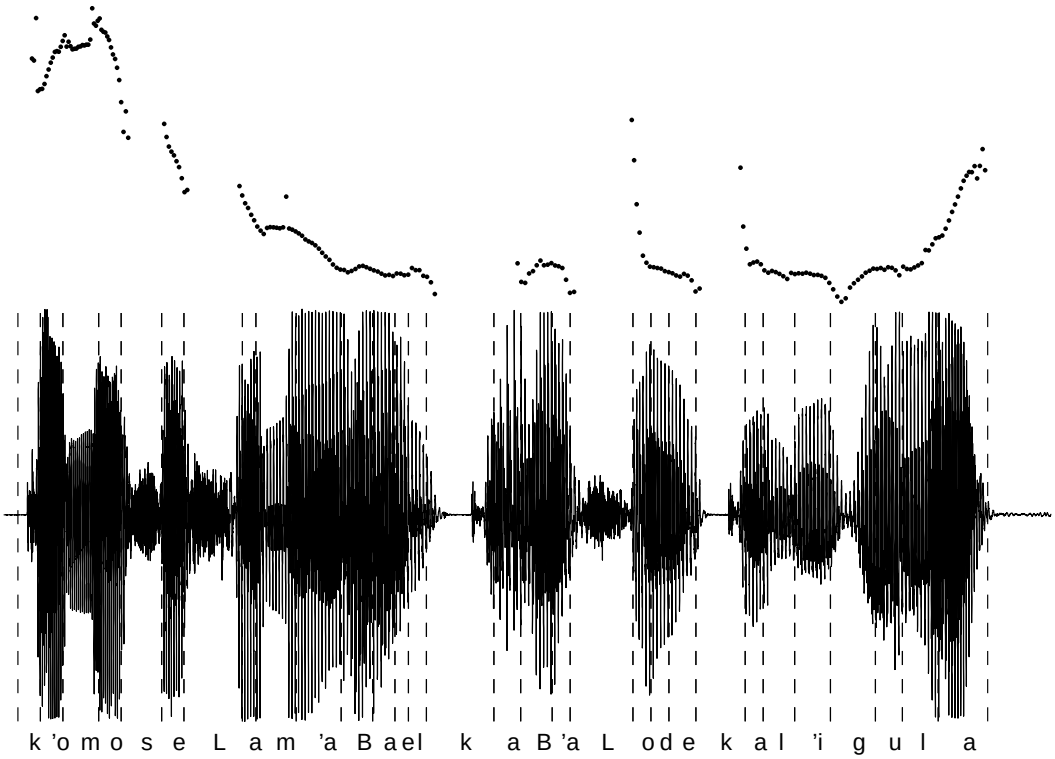


Figura 2.2: Ejemplo de contorno de entonación: *¿Cómo se llamaba el caballo de Calígula?*

en los segmentos sordos debido a la ausencia de vibración de las cuerdas vocales que imposibilitan la medición de dicho parámetro.

En la literatura existe una gran variedad de algoritmos de extracción para estimar la curva de frecuencia fundamental. Estos usan diversas representaciones de la señal del habla: en el dominio del tiempo ([Phi85, Sec83, Med91]), de la frecuencia ([Sch68, Nol70]) o de la *quefrency* (cepstrums) ([Nol67]). La diversidad y la complejidad de los métodos usados para determinar la frecuencia fundamental instantánea derivan del carácter no estacionario del habla, caracterizado por una intensidad no uniforme, pequeñas perturbaciones en la frecuencia fundamental entre los sucesivos periodos debido al tipo de fonación, y por un cambio constante del espectro ocasionado por la articulación.

Una vez obtenida la curva de entonación, es necesario usar diferentes unidades para su análisis. Cada una de ellas tiene un diferente alcance en términos de duración y permite estudiar distintos aspectos de la misma. En la siguiente sección describiremos las unidades de la entonación, para luego dedicarnos al modelado de la entonación en los sistemas de conversión texto a voz (Sección 2.2.2).

2.2.1. Unidades de la entonación

La entonación se puede analizar usando un conjunto de unidades entonativas organizadas jerárquicamente que abarcan desde el fonema hasta varias oraciones. Los diferentes alcances de cada unidad permiten hacer un análisis detallado de los fenómenos locales y globales. Las unidades de entonación más pequeñas se centran en los fenómenos locales mientras que las unidades más largas son usadas en el estudio de fenómenos globales.

Este punto de vista jerárquico de las unidades de entonación es parte de la descripción utilizada por el enfoque superposicional, donde es posible ver a la entonación como el resultado de la composición de los comportamientos de estas unidades, cada cual con su alcance limitado.

La sílaba se puede considerar como la unidad de entonación más pequeña, con la cual es posible analizar el fenómeno del acento léxico. El acento se ve reflejado como una subida o bajada de la entonación para enfatizar una sílaba con respecto a las sílabas vecinas, tanto acentuadas como no acentuadas. Tal diferenciación es necesaria para propósitos léxicos y rítmicos. Por ejemplo, en muchos idiomas el significado de la palabra depende de la sílaba acentuada.

Además, en este nivel también es posible estudiar el grado de cambio del tono, su temporización y forma, y la influencia en los sonidos vecinos. El estudio de estos aspectos es necesario para la comprensión de la entonación, para su posterior utilización en el modelado de la prosodia en un conversor texto a voz.

La siguiente unidad en jerarquía es el grupo acentual. Es la unidad más pequeña con significado porque abarca una o más palabras, y su alcance depende del idioma. En algunos de ellos, como en el inglés, el grupo acentual se define como la sílaba acentuada y todas las siguientes hasta la próxima sílaba acentuada [Spr98a].

En el caso del español, para algunos autores el grupo acentual se encuentra constituido por una palabra acentuada y todas aquellas palabras no acentuadas que le preceden. Esta unidad ha sido utilizada por numerosos autores para describir los patrones de entonación del español a nivel local [Gar96, Alc98, Sos99, Esc02a].

El grupo acentual incluye información sobre el acento, y se distinguen tres tipos principales de grupos acentuales en función de la posición de la sílaba acentuada en la última palabra: agudos (última sílaba acentuada), graves (penúltima sílaba acentuada) y esdrújulos (antepenúltima sílaba acentuada). Sin embargo, debido a la aparición de algunos sufijos (por ejemplo: -mente), existen más tipos de grupos acentuales debido a configuraciones más complejas de sílabas acentuadas.

Generalmente, los grupos acentuales se combinan para formar unidades de entonación más grandes, tales como el grupo entonativo y la cláusula entonativa [Gil04]. Estas unidades se pueden identificar por los eventos prosódicos que ocurren en sus fronteras y que determinan los límites de la unidad y su clasificación. Estos eventos son discontinuidades en el contorno entre diferentes secciones de la oración, pausas, aumento de la duración de la sílaba final y una disminución de la velocidad del habla. Dependiendo de las características de estos eventos prosódicos, la unidad se clasifica como una frontera de grupo o de cláusula entonativa.

Un grupo entonativo presenta en su frontera pausas o inflexiones de la frecuencia fundamental, y es el ámbito en el que se han definido habitualmente los patrones melódicos [Qui93]. Los grupos melódicos se agrupan dentro de la cláusula entonativa, que se define como un conjunto de grupos entonativos afectados por el mismo patrón de supradeclinación [Gar01].

Estas unidades están más relacionadas con aspectos amplios del discurso: una subida del tono en un grupo entonativo indica continuación antes de una coma, o bien el tipo de movimiento tonal al final de oración diferencia una frase interrogativa de una declarativa.

Uno de los aspectos importantes desde el punto de vista de los ordenadores aplicados a TTS es que los grupos acentuales y las sílabas son unidades de entonación que pueden ser delimitadas usando la información en el texto. Por lo tanto, sus fronteras se pueden determinar mediante algoritmos automáticos. Sin embargo, los grupos y cláusulas entonativas dependen mucho de la decisión del locutor y del significado de la oración, existiendo múltiples configuraciones válidas. La mala ubicación de una juntura terminal puede cambiar sustancialmente el significado de una oración. En los TTS existen módulos dedicados a la predicción de las fronteras de grupos y cláusulas entonativas que poseen una precisión limitada debido a las razones mencionadas anteriormente.

2.2.2. La entonación en la conversión texto-voz

En los conversores texto a voz, la información extraída del texto de entrada es usada para generar un contorno adecuado de entonación. Dicha información es provista por el módulo de procesamiento de texto: normalización, separación en sílabas, acento léxico, información morfológica, etc.

Después de procesar esa información, se obtienen un conjunto de características (F) que serán usadas para tomar la decisión sobre el contorno de entonación más adecuado. Para ello, una función de mapeo relaciona el espacio de características (F) con el espacio de contornos de entonación (f_0). Esta función constituye el modelo de entonación en un TTS.

$$G(F) = f_0 \tag{2.3}$$

La función de mapeo es solamente una aproximación al contorno de entonación real (f_0) que realizaría un hablante. Siempre existe un error e ($f_0 = G(F) + e$) debido al ruido en el contorno de frecuencia fundamental (errores en los algoritmos de extracción y la microprosodia), limitaciones en las características disponibles (algunas características faltantes reducen la dimensión del espacio F y provocan que sea imposible encontrar algunas interrelaciones debido al solapamiento de sus efectos) y decisiones particulares por parte del hablante que son imposibles de predecir.

En las siguientes secciones se describirán distintos enfoques utilizados en el modelado de la entonación, con aplicaciones para la conversión texto a voz. La clasificación utilizada es similar a la propuesta por Botinis et al. [Bot01]: modelos fonológicos, modelos perceptuales y modelos de estilización acústica superposicionales y no superposicionales.

2.2.3. Modelos de entonación fonológicos

Estos modelos describen a la entonación como una secuencia de tonos asociados a diferentes eventos del contorno de frecuencia fundamental. Para ello se emplea un conjunto reducido de símbolos para representar los distintos tonos.

La capacidad representativa de estos modelos está limitada por el número de símbolos y las restricciones para las combinaciones de los mismos, a través de gramáticas particulares para cada idioma.

Uno de los modelos que utiliza este tipo de enfoque es el propuesto por Pierrehumbert [Pie80]: descripción de la entonación autosegmental métrica. El mismo describe la entonación del inglés empleando dos categorías de tonos: alto (H) y bajo (L). Los tonos no interactúan, sino que simplemente se componen secuencialmente en el tiempo.

Este modelo contiene unos símbolos para indicar tonos de frontera de la cláusula entonativa: H % y L %, ya que los mismos están asociados al final del grupo de entonación más marcado acústica y perceptualmente. Los acentos tonales poseen varios símbolos representativos: H*, L*, L*+H, H*+L, L+H* y H+L*. El tono con asterisco (por ejemplo, **L***+H) está asociado con la sílaba acentuada, mientras que el otro se asocia a las sílabas que preceden o siguen a la acentuada (por ejemplo, L*+**H**). El acento de frase (H- o L-) indica el tono de frontera para los grupos entonativos.

Las reglas del modelo de Pierrehumbert son la base del sistema de transcripción prosódica ToBI (*Tones and Break Indices*) presentado por Silverman [Sil92]. Con el fin de instruir a los transcriptores de ToBI existe un documento llamado "Guidelines for ToBI Labelling", que incluye una serie de ejercicios de práctica. Es importante remarcar que existen experimentos demostrando un alto grado de acuerdo entre transcriptores diferentes [Pit94], gracias a dicha instrucción. Sin embargo, en su artículo Wightman [Wig02] observa que este alto grado de acuerdo solamente ocurre para un subconjunto de las etiquetas. Además, el mismo autor observa que el etiquetado es muy lento, pudiendo resultar de 100 a 200 veces el tiempo real [Syr01].

La aplicación de este modelo a la conversión texto a voz requiere la definición de puntos objetivo tanto en los acentos tonales como en los tonos de frontera. Los puntos objetivo se estiman obteniendo tanto la amplitud como el instante de tiempo de los mismos, ya sea aplicando un sistema de reglas [And84, Möh95] o métodos estadísticos basados en regresión lineal [Bla96].

Otro sistema de codificación descrito en la literatura es INTSINT [Hir94]. En esta codificación también se señalan los eventos significantes de la curva tonal usando un conjunto limitado de símbolos para señalar tonos absolutos (T, M y B) y relativos (H, L, S, U y D).

Los tonos absolutos en INTSINT se definen de acuerdo al rango tonal del hablante, mientras que los relativos se anotan con respecto a la altura tonal de los puntos adyacentes. En su conjunto permiten hacer una descripción detallada del contorno de frecuencia fundamental a través del análisis automático de la entonación [Hir00] usando una herramienta de estilización de contornos: MOMEL [Hir93].

Uno de los aspectos remarcables de INTSINT es que la transcripción conserva los

valores numéricos de los eventos tonales. Por lo tanto, es posible representar la curva tanto en forma cualitativa (como en el caso de ToBI) como cuantitativa (parametrizada). Las correlaciones lingüistico/funcionales de estos eventos pueden vincularse con un análisis de las propiedades pragmáticas, semánticas y sintácticas de la oración.

2.2.4. Modelos de entonación perceptuales

Los modelos perceptuales se basan en el hecho de que solamente algunos movimientos tonales son perceptibles. Debido a esto, solo dichos movimientos deberían ser estudiados para modelar la entonación.

El esquema IPO [Har90] es el modelo perceptual más conocido. En dicho modelo el contorno de frecuencia fundamental es estilizado utilizando segmentos rectos, creando una versión perceptualmente equivalente más sencilla. Luego, los patrones extraídos con la representación rectilínea son caracterizados tanto en duración como en amplitud. Finalmente, se crea una gramática que limita el número y tipo de movimientos permitidos para una lengua dada.

D'Alessandro y Mertens [d'A95, Mer97] propusieron un método para automatizar el proceso de estilización. Para ello analizan la parte vocálica de cada sílaba incluyendo un vértice cuando la distancia de las rectas que lo definen con respecto al perfil original supera un umbral de percepción. Aquí cada sílaba puede ser modelada por más de un segmento, a diferencia del modelo IPO donde un segmento puede implicar más de una sílaba.

2.2.5. Modelos de entonación por estilización acústica superposicionales y no superposicionales

En la literatura se han propuesto muchos modelos de entonación por estilización acústica superposicionales y no superposicionales. Sin embargo, en general se puede afirmar que la generación de los mismos consiste en cuatro pasos básicos: selección de la unidad/es de entonación, parametrización de los contornos de entonación, extracción de las características de cada unidad y estimación de la función de mapeo $F \rightarrow f_0$. En las siguientes secciones se estudiarán cada uno de ellos.

Selección de la unidad/es de entonación

La elección de la unidad de entonación es esencial porque condicionará los dos pasos siguientes: parametrización y extracción de características. Tanto la sílaba como el grupo acentual son las unidades de entonación más comúnmente usadas. Ambas se enfocan en el acento como el punto clave para el modelado de la entonación.

Por otra parte, las fronteras prosódicas como el grupo y la cláusula entonativa son modeladas en algunas ocasiones usando un enfoque superposicional, como es el caso del modelo de Fujisaki [Fuj84]. En otros enfoques, algunos autores como Escudero [Esc02b], modelan el acento y las fronteras prosódicas con la misma unidad de entonación y el mismo conjunto de parámetros.

Parametrización de los contornos de entonación que abarca cada unidad

La representación paramétrica es una aproximación del contorno real de la unidad entonativa, la cual permite un análisis más compacto de la trayectoria tonal. La función de aproximación puede corresponder a muchas formulaciones matemáticas (exponenciales, polinomiales, etc.), cuyos parámetros deben ser estimados calculando su ajuste con los contornos de entonación reales. En general, tales parámetros deben satisfacer un conjunto de requisitos:

- **Representativa.** Los parámetros deben describir la forma del contorno de entonación de una manera significativa. A través de esto será más fácil el análisis de las relaciones entre los parámetros y las características.
- **Homogénea.** Es preferible que el número de parámetros sea el mismo para aproximar cualquier contorno de entonación. Tal homogeneidad facilitará el uso de algoritmos de agrupamiento y a los métodos de aproximación. El uso de diferentes conjuntos de parámetros para cada caso incrementa la complejidad del modelo.
- **Precisa.** La precisión del ajuste de la representación paramétrica se debe adecuar a la tarea. Algunos dominios tienen contornos de entonación que son curvas simples y suaves (por ejemplo: habla neutra). En otras situaciones los contornos de entonación pueden ser lo suficientemente complejos para tener muchas fluctuaciones dentro de un fonema (por ejemplo: lectura de cuentos para niños). En consecuencia, es necesario analizar la complejidad de la entonación para alcanzar una precisión adecuada.
- **Estimable.** La estimación de los parámetros en base a los contornos de entonación debe ser posible. Dicha estimación puede ser realizada usando diferentes herramientas matemáticas, tales como álgebra matricial, métodos de gradiente o algoritmos genéticos. Este paso es puramente matemático y pertenece al campo de ajuste de curvas paramétricas a un conjunto de datos. Tales ajustes permitirán el tratamiento de contornos de entonación de diferente duración y eventos tonales localizados en diferentes posiciones. Además, es aconsejable un modelo de entonación que permita una solución cerrada para la estimación de los parámetros frente a modelos que exijan el uso de métodos de gradientes para encontrar la solución, debido a razones de exactitud y de tiempo de cálculo.
- **Capacidad de generalización.** Uno de los principales objetivos de los modelos de entonación es predecir contornos de entonación para conjuntos de valores de las características no observados en el momento del entrenamiento.

Por lo tanto podemos concluir que la elección de la formulación matemática constituye una parte importante en la obtención del modelo de entonación. La elección debería obedecer al principio de “la Navaja de Occam”: mantener la complejidad lo más pequeña posible [RA99].

Extracción de las características de cada unidad

Las características F del modelo de entonación $f_0 = G(F)$ suelen ser de carácter morfológico y sintáctico, y muchas de las utilizadas en esta tesis han sido propuestas por diversos autores, tales como Lopez [Lóp93], Garrido [Gar96], Alcoba [Alc98] y Vallejo [Val98]. La extracción de estas características se realiza usando la información provista por el módulo de procesamiento de texto: normalización, separación en sílabas, acento léxico, información morfológica, etc.

Las características principales relacionadas con la sílaba son el acento léxico y su posición en la palabra, que determinarán el nivel de variación del contorno de la sílaba y su dirección. La posición de la sílaba dentro de la palabra determinará también la forma de variación del contorno, por ejemplo, dependiendo de su proximidad a otra sílaba acentuada.

Dentro de las características de la palabra se encuentra la posición del acento léxico, el énfasis, la posición en el grupo acentual y entonativo, la posición en la cláusula entonativa y la información morfosintáctica.

La posición del acento léxico determinará el instante de la excursión en el contorno de frecuencia fundamental para indicar la sílaba acentuada. Esto es de importancia, ya que idiomas como el español distinguen el significado y la función de la palabra según la sílaba acentuada. Por ejemplo, de**pó**sito (sustantivo), depo**si**to (verbo presente) y deposi**tó** (verbo pasado). Por otra parte, el énfasis también influye en el grado de variación del contorno de frecuencia fundamental, y de esta manera el locutor puede indicar diferentes niveles del mismo.

La posición dentro del grupo acentual en el grupo entonativo y en la cláusula entonativa tiene una gran influencia en la forma del contorno de frecuencia fundamental. Los grupos acentuales más característicos son los iniciales y finales. Los iniciales suelen presentar una subida constante hasta el final de la sílaba acentuada o la sílaba siguiente. Los intermedios presentan en general una caída inicial más una subida constante hasta el final de la sílaba acentuada o la siguiente. El grupo acentual final incluye la juntura terminal y el tipo de oración.

El tipo de frase es uno de los aspectos que influye de manera más clara en la forma de la entonación, conjuntamente con el acento léxico. Si se considera la parte final del grupo entonativo o la cláusula entonativa, se distinguen junturas terminales con contorno descendente correspondientes a frases enunciativas y contorno ascendente correspondientes a frases interrogativas. Las frases exclamativas determinan la forma del contorno de frecuencia fundamental a lo largo de toda la frase, y no solamente en la región final, como es el caso de las oraciones declarativas e interrogativas.

Otra característica importante que influye en la entonación son los signos de puntuación. Algunos de ellos se utilizan como indicadores de tipo de frase: signo de exclamación e interrogación, y el punto. Otros signos de puntuación poseen contornos de entonación particulares, tal como ocurre con la coma y los dos puntos, que en ocasiones es encuentran asociados con contornos que indican una continuación.

El tipo de discurso y el dominio son características importantes en el modelado de la entonación. El dominio se modela en muchas ocasiones utilizando datos específicos

para el mismo, sin introducir otros. Sin embargo, tanto los elementos discursivos dentro del dominio como la existencia de sub-dominios dentro del mismo pueden introducir una variabilidad que debería ser modelada. Por ejemplo, un modelo de entonación aplicado a la lectura de noticias debería distinguir entre diferentes sub-dominios, tales como el de las noticias deportivas y las noticias internacionales. Las noticias deportivas se leen en ocasiones con un mayor grado de exclamación y excitación. Mientras tanto, noticias internacionales referidas a acontecimientos graves son leídas con mayor seriedad y respeto.

El estado de ánimo o la intención son también elementos importantes que determinan la forma del contorno de frecuencia fundamental. Sin embargo, son difíciles de extraer con las técnicas de procesamiento del lenguaje natural de hoy en día. Lo mismo ocurre con otras muchas características sintácticas, semánticas y pragmáticas.

Estimación de la función de mapeo $F \rightarrow f_0$

La función de mapeo $F \rightarrow f_0$ relaciona las características extraídas de la unidad con la representación paramétrica de la misma. De esta manera, se intenta obtener el contorno de frecuencia fundamental de la unidad con el menor error de estimación posible con respecto al contorno real, dadas las características lingüísticas extraídas de la misma. En la etapa de síntesis o generación esta función nos generará el contorno sintético $G(F)$.

En la literatura existe una gran variedad de modelos de entonación, con diversas funciones matemáticas y métodos de entrenamiento, que usan enfoques de aprendizaje automático basados en datos en la mayoría de los casos. Las técnicas de aprendizaje automático se aplican a los datos para extraer regularidades en la relación entre las características lingüísticas del mensaje y el comportamiento del contorno de frecuencia fundamental. Son ampliamente utilizadas debido a varias ventajas que presentan:

- **Entrenamiento automático.** El análisis y la generación de reglas manuales que expliquen el comportamiento del contorno de frecuencia fundamental requiere de personas entrenadas y un largo tiempo de desarrollo. Es preferible usar técnicas automáticas con poca supervisión que puede encontrar regularidades en grandes volúmenes de datos de entrenamiento.
- **Rápida adaptación a nuevos dominios usando datos adecuados.** El tiempo de adaptación a nuevos dominios es importante debido a que eso puede condicionar el tiempo de desarrollo de un proyecto. Las técnicas de aprendizaje automático pueden extrapolar el conocimiento adquirido en un dominio a otro. De esta manera, se evita el desarrollo desde cero de un sistema y se acelera el proceso de migración de dominio.
- **Uso de características continuas y discretas.** Las características usadas en los conversores texto a voz son tanto continuas (por ejemplo: duración) como discretas (por ejemplo: etiquetas morfológicas). Las técnicas de aprendizaje automático puede proporcionar nuevo conocimiento acerca de la tarea analizando las reglas obtenidas luego del entrenamiento.
- **Casos no observados**. Las técnicas de aprendizaje automático pueden encontrar una estructura dentro de los datos y extrapolarla a casos no observados. Sin embargo,

aparecen limitaciones debido a características faltantes en el texto de entrada que no pueden ser obtenidas debido a las limitaciones de entendimiento de las computadoras.

Entre los algoritmos de aprendizaje automático por computadora se pueden mencionar: árboles de clasificación y regresión (CART) [Bre84], redes neuronales (NN) [McC43], aprendizaje basado en la memoria (MBL) [Sta86], etc.

En este punto es necesario aclarar que una de las desventajas de las técnicas de aprendizaje automático son las medidas objetivas usadas para valorar el proceso de entrenamiento. En algunos casos estas no están relacionadas íntimamente con la psicoacústica debido a dificultades de implementación. En general, estas miden la correlación o el error cuadrático medio entre los contornos de referencia y los predichos. Estas medidas globales no se enfocan en ciertos aspectos locales que pueden producir una opinión más baja en los evaluadores (valoración subjetiva). Por ejemplo: un desplazamiento en el acento en una palabra puede contribuir a un MOS (Mean Opinion Score) más bajo. Por tanto, el criterio que guía los métodos automáticos puede afectar negativamente a sus prestaciones.

El modelo de Fujisaki

Fujisaki [Fuj84] desarrolló un modelo matemático del proceso de generación del contorno de F_0. A pesar que el modelo fue desarrollado inicialmente para el japonés, el mismo puede ser aplicado a muchos otros idiomas mediante algunas modificaciones específicas para cada uno de ellos [Fuj98].

Este modelo tiene el sustento de una justificación fisiológica para la formulación matemática basado en observaciones fisiológicas de la dinámica de la laringe [Fuj00b]. Básicamente, Fujisaki derivó un modelo sobre la influencia de la tensión y la elongación de los músculos esqueléticos en la frecuencia de vibración de una membrana elástica.

El modelo de entonación de Fujisaki modela el contorno entonativo como la suma de la salida de dos filtros de segundo orden (Figura 2.3). El filtro de la primera rama es excitado por impulsos denominados comandos de frase (Ecuación 2.5) de amplitud A_p y ubicación temporal T_0. La respuesta impulsional del filtro es larga y modela la declinación o evolución del grupo entonativo. El segundo filtro se encuentra excitado por pulsos denominados comandos de acento (Ecuación 2.6) de amplitud A_a y ubicación temporal T_1 y T_2. La respuesta de estos filtros es más limitada en el tiempo, y permite modelar las variaciones de F_0 más localizadas, tales como las que ocurren en una sílaba o un grupo acentual. Una componente de continua (F_b) se suma a la salida de estos filtros (Ecuación 2.4), que contribuye a ajustar el valor inferior del rango de la entonación.

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{p_i} G_p(t - T_{0_i}) + \sum_{j=1}^{J} A_{a_j} G_a(t - T_{1_j}) - G_a(t - T_{2_j}) \tag{2.4}$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{2.5}$$

$$G_a(t) = \begin{cases} \text{mín}[1-(1+\beta t)e^{-\beta t}, \gamma] & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{2.6}$$

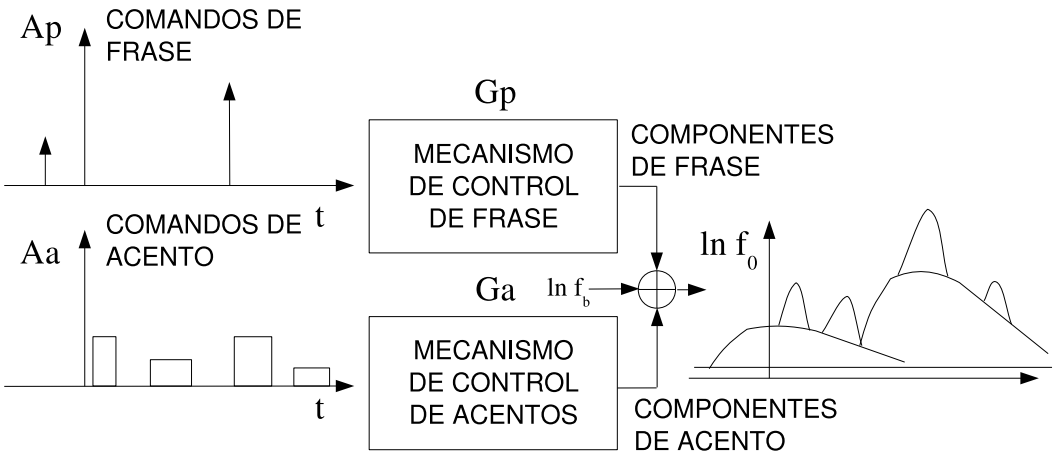


Figura 2.3: Esquema del modelo de entonación de Fujisaki.

Su principal desventaja es que no es posible obtener una solución cerrada que encuentre los parámetros A, T, F_b, i y j a partir de un contorno. Por lo tanto, es necesario hacer uso de técnicas de optimización basadas en gradientes para obtener la solución.

En esta tesis veremos que es posible alcanzar una solución cerrada para la amplitud de los comandos si se asume que los instantes de tiempo son conocidos (T_0, T_1 and T_2), tal como se demuestra en la Sección 3.1.4 [Sil04, Agü05], lo que facilita la búsqueda de la solución.

Mixdoff propuso un algoritmo de extracción de parámetros para el modelo de entonación de Fujisaki usando múltiples componentes del contorno de frecuencia fundamental [Mix00]. El primer paso del algoritmo consiste en un suavizado usando el algoritmo MOMEL [Hir00], que es una estilización cuadrática spline de la curva. Luego, se realiza una descomposición del contorno usando filtros paso-altos y paso-bajos. La salida del filtro paso-altos tiene en cuenta los movimientos rápidos del contorno, y en esta componente son detectados los comandos de acento. Debido a que el filtro paso-bajos contiene las variaciones más lentas del contorno, los comandos de frase se detectan en esta componente. Una vez detectados los comandos de frase y acento, se aplica un algoritmo de optimización para refinar tanto las amplitudes como los instantes de tiempo, minimizando el error cuadrático medio de la predicción.

Fujisaki et al. [Fuj00a, Nar02b, Nar02a] sugieren un algoritmo ligeramente diferente al propuesto por Mixdorff. La extracción de parámetros comienza aplicando un preprocesamiento que resulta en una estilización usando polinomios de tercer orden continuos. El primer procesamiento elimina los errores grandes y los efectos de borde del contorno original. El contorno resultante es una interpolación por partes cúbica, la cual resulta en una curva diferenciable en todos los instantes. Los comandos se buscan en la derivada de esta función. Una secuencia de máximo y mínimo en la primer derivada corresponde

a las fronteras de un comando de acento. Una vez extraídos los comandos de acento, los comando de frase se detectan en el residuo resultante.

Los algoritmos explicados anteriormente realizan suposiciones que pueden afectar la estimación de los parámetros, tales como las técnicas de suavizado para asegurar continuidad y derivabilidad de los contornos, y el uso de filtros o derivadas para la detección de los comandos. Por otra parte, el número de comandos y sus valores depende altamente de la estilización. Como consecuencia, los parámetros no estarán necesariamente relacionados con el fenómeno fisiológico.

Finalmente, los comandos no podrán ser predichos directamente usando el contenido lingüístico, debido a que no se ha usado dicha información para condicionar los comandos extraídos. Algunos métodos propuestos en la literatura abordan este problema presentando soluciones a través del uso de limitaciones lingüísticas durante el proceso de extracción de parámetros [Möb95, Nav02b, Hir03, Agü04b].

El modelo Tilt

Taylor [Tay00] propuso un modelo de entonación caracterizado por una secuencia de eventos de entonación: acento y tonos de frontera. Cada evento tiene una componente de ataque, una de decaimiento, o ambas.

La parametrización se puede observar en la Figura 2.4. La amplitud y la duración de los parámetros describe la trayectoria de la curva de entonación. Sin embargo, estos parámetros son relativos a los otros dos parámetros $F0_{peak}$ y t_{peak}, que pueden ser tanto absolutos como relativos. $F0_{peak}$ depende del rango tonal del hablante mientras que t_{peak} es relativo al tiempo de comienzo del núcleo silábico.

El modelo Tilt es muy conocido debido a que se emplea en Festival, un conversor texto a voz de distribución gratuita ampliamente utilizado, que es desarrollado por el Centre for Speech Technology Research CSTR de la Universidad de Edimburgo.

Para la detección automática de los eventos se han desarrollado algoritmos que utilizan Modelos Ocultos de Markov [Tay93]. Una vez detectados, la extracción de los parámetros puede hacerse automáticamente empleando el análisis RFC (Rise-Fall-Connection: Ataque-Decaimiento-Conexión) [Wri97]. Tanto los acentos como las fronteras entonativas se describen usando elementos de ataque y decaimiento, y la unión entre ellos se realiza mediante los elementos de conexión.

Para la predicción de los parámetros del modelo Tilt para la síntesis de contornos de frecuencia fundamental se ha propuesto la utilización de árboles binarios de decisión y regresión, como se puede observar en la tesis de Dusterhoff [Dus00] y en el software Festival.

Modelado de la entonación basado en curvas de Bézier

En su tesis [Esc02b], Escudero propone el uso de curvas de Bézier para el modelado de la entonación, las cuales se basan en una función polinómica. Los coeficientes de

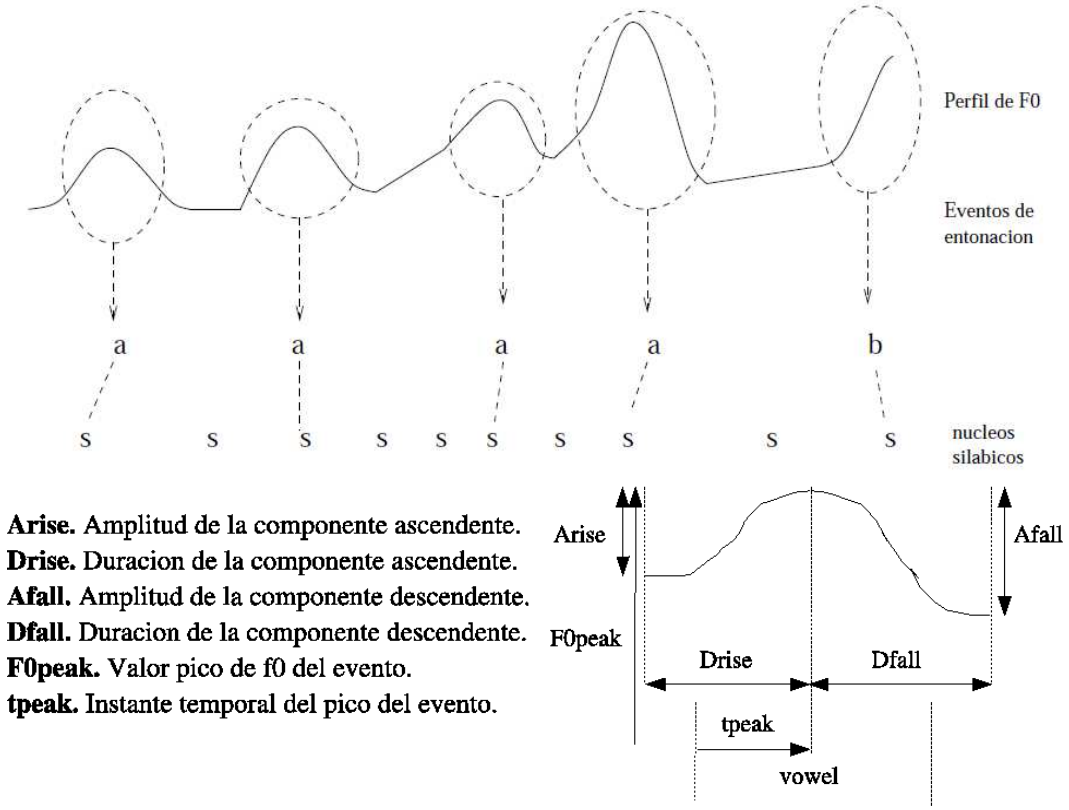


Figura 2.4: Parámetros Tilt.

Bézier permiten una representación más significativa que la resultante de los coeficientes polinómicos en forma expandida.

La formulación polinómica se detalla en la ecuación 2.7 y la forma de los polinomios base para una curva de cuarto orden se encuentran en la Figura 2.5. Como vemos, un polinomio de orden N se representa como N polinomios base.

$$P(t) = \sum_{n=0}^{N} \alpha_n \begin{pmatrix} N \\ n \end{pmatrix} t^n \, (1-t)^{(N-n)} \tag{2.7}$$

En su tesis, Escudero representa cada grupo acentual con un polinomio de tercer grado. Analiza varias maneras de clasificar grupos acentuales basándose en las propuestas de Lopez [Lóp93], Garrido [Gar96], Vallejo [Val98] y Alcoba [Alc98]. De esta manera, un contorno de frecuencia fundamental se puede predecir usando las características lingüísticas del grupo acentual. En la Figura 2.6 se muestra el uso de las curvas de Bézier para grupos acentuales, con restricciones de continuidad hasta la primera derivada. El objetivo de estas restricciones es tener en cuenta durante la aproximación el contexto en el que se realizan los grupos acentuales correspondientes.

Otro modelo de entonación similar propuesto en la literatura es el descrito por Veronis

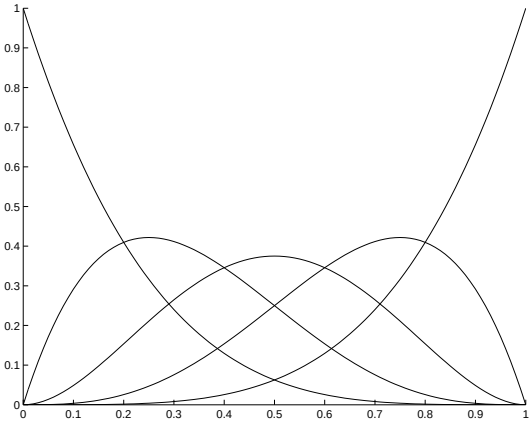


Figura 2.5: Polinomios de Bézier de orden cuatro

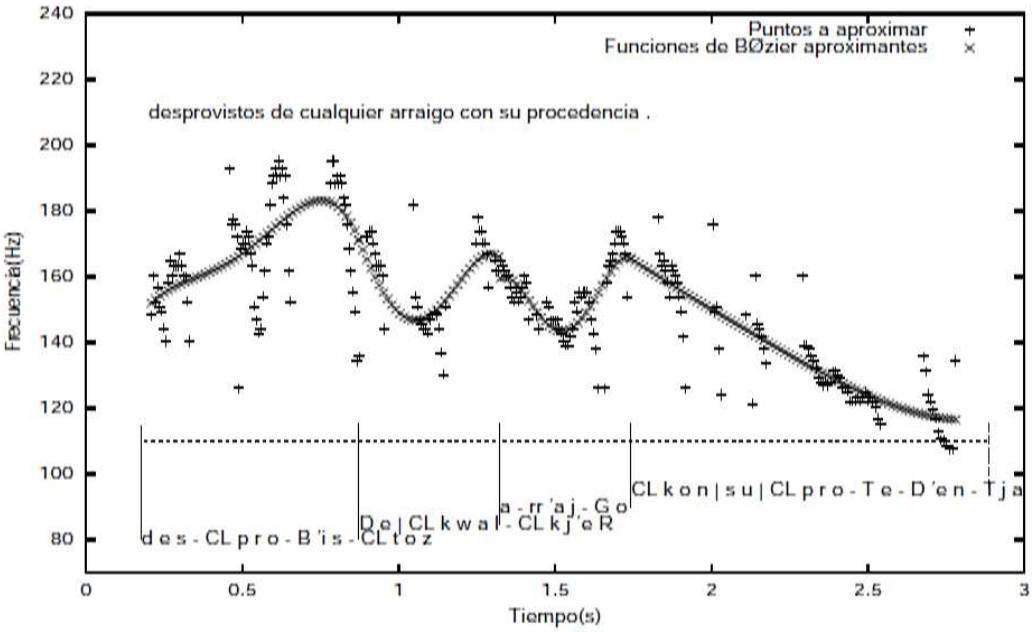


Figura 2.6: Contorno de frecuencia fundamental aproximado usando curvas de Bézier con cinco coeficientes [Esc02b].

et al [Vér98], en el cual se utilizan splines cuadráticos correspondientes a la estilización MOMEL [Hir93]. La entonación es predicha utilizando etiquetas prosódicas abstractas correspondientes a INTSINT [Hir94], que son deducidas a partir de la información gramatical contenida en el texto.

2.3. Duración

La duración de los segmentos es otro importante parámetro prosódico en la conversión texto a habla, ya que transporta información tanto sobre el hablante como sobre el mensaje, debido a que está íntimamente relacionada con la percepción del ritmo o cadencia. La cadencia es esencial en la comunicación, debido a que se usa para expresar muchos aspectos, tales como los acentos, separar información en constituyentes, y para la articulación natural de los sonidos. Además es un indicador de énfasis y del estado de ánimo. Es muy común usar una cadencia lenta para mostrar un estado depresivo.

La duración segmental se define como el intervalo de tiempo entre las fronteras de la unidad segmental: el fonema. A continuación se describirán diferentes factores que influyen en la variación de la duración de los fonemas, para luego abordar el proceso de generación de la duración en la conversión texto a voz.

2.3.1. Factores que influyen en la variación de la duración segmental

En el estudio de la duración segmental existen diversas fuentes de variación. Estas fuentes son clasificadas por Klatt [Kla76] en factores segmentales, silábicos y suprasilábicos, que serán discutidos en los siguientes párrafos, conjuntamente con otro importante factor: la velocidad del habla [Whi02].

Los **factores segmentales** de la duración son aquellos que son intrínsecos a los fonemas o a las características de los mismos, como por ejemplo, las diferencias entre vocales cortas y largas, la mayor duración de las fricativas sordas que las fricativas sonoras, la mayor duración de las oclusivas bilabiales que las alveolares o velares, etc. En general, estos factores son explicaciones articulatorias de las variaciones de la duración. Por ejemplo, las vocales cortas son más centrales y por lo tanto su articulación requiere menos tiempo. Otro ejemplo son las oclusivas bilabiales que generan una cavidad supralaríngea más grande que las oclusivas alveolares o velares. En consecuencia, necesitan más tiempo para producir la suficiente presión en el punto de constricción para generar una liberación de presión audible.

A pesar de que las consideraciones articulatorias son importantes, la fonología del idioma juega un rol clave en la explicación de tales variaciones. Por ejemplo, en noruego existe una clara distinción entre las vocales cortas y largas: las vocales largas deben ser marcadamente largas, y las vocales cortas a veces son omitidas.

Otra fuente de variación de la duración son los **factores silábicos**, que resultan de la organización de los segmentos en una cadena de sílabas, las cuales son consideradas generalmente los elementos constituitivos de las palabras.

Uno de los factores silábicos más destacados es el aumento de la duración de las vocales en sílabas acentuadas (las consonantes en el ataque son también más largas en las sílabas acentuadas). La diferencia en la duración tiene su origen en factores lingüísticos: se usa una duración mayor en las sílabas acentuadas determinado por la morfología y no por motivos articulatorios.

Los **factores suprasilábicos** que afectan a la duración provienen de la estructura

lingüística de la cadena de sílabas. Estos pueden ser clasificados en tres categorías: aumento de la duración debido a fronteras de unidades entonativas o a la prominencia, y disminución de la duración debido al tamaño del contenido fonológico de los constituyentes.

La **temporización** de un idioma es otro factor silábico importante. Existen tres tipos básicos de temporización: acento (por ejemplo: el inglés), sílaba (por ejemplo: el español) y mora (por ejemplo: el japonés). La influencia del tipo de temporización determina como debe ser la duración de las sílabas para preservar la temporización del idioma.

Además de los factores mencionados anteriormente, los cambios en la **velocidad del habla** puede estar motivados por factores lingüísticos y no-lingüísticos, y esto ejerce una influencia en la duración de los segmentos. Cuando los hablantes reducen la velocidad de locución, una importante fracción de la duración extra es introducida a través de pausas [GE68] para evitar duraciones segmentales raras y articulaciones dificultosas. En cambio, el aumento de la velocidad se encuentra acompañada por simplificaciones fonéticas.

El término velocidad del habla se puede usar con dos sentidos. Más comúnmente es usado para significar la velocidad global del habla, es decir, la duración de los segmentos del habla en una determinada estructura (grupo entonativo, oración, párrafo) debido a factores tales como fisiología, dialecto, emociones, etc. En general se mide en palabras o sílabas por segundo.

En otros casos este término se refiere a la velocidad local del habla debido a la influencia de la estructura lingüística en la temporización del habla. Esta última es más difícil de medir y ha sido motivo de estudio en varias publicaciones [Pfi96, Wan07].

2.3.2. Generación de la duración en los TTS

El modelo de duración en los conversores texto a voz tiene como objetivo la generación de una duración adecuada para cada segmento. Esta tarea se realiza usando información provista por el módulo de procesamiento de texto: estructura del texto, silabificación, acento léxico, etiquetas morfosintácticas, transcripción fonética, características articulatorias de los fonemas, etc.

Dicha información se usa para generar las características F, las cuales están íntimamente relacionadas con los factores causantes de la variación de la duración explicados en la sección anterior. Dichas características serán usadas para estudiar su relación con la duración de los segmentos D, con el objetivo de obtener una función G que genere una duración D dadas las características F:

$$G(F) = D \tag{2.8}$$

Esta función es solamente una aproximación. La formulación real tiene un error de aproximación e que debe ser minimizado.

$$G(F) = D + e \tag{2.9}$$

En resumen, para realizar el estudio de la influencia cuantitativa de las características

F en la duración de los fonemas es necesario disponer de audios con segmentación fonética (manual o automática) y de un conjunto de características relacionadas con su variación.

Los mejores resultados para la síntesis se obtienen usando segmentación manual de los datos. Sin embargo, algunos investigadores (como por ejemplo [Mak00]) indican que los métodos automáticos para la segmentación del habla puede alcanzar buenos resultados para su uso en lugar de métodos manuales. Este enfoque permite ahorrar tiempo de desarrollo y disminuir costos.

Para evaluar los modelos de duración es importante disponer de fronteras de referencia de calidad. Los errores de los sistemas automáticos son mayores a los 10ms para el $10-20\,\%$ de los casos [DTT03, Ade05]. Esta magnitud del error de segmentación es importante en comparación con el RMSE de los modelos de predicción de la duración en TTS, y deben ser considerados en la evaluación de la calidad de los modelos.

El conjunto de características extraídas del texto se usa para obtener la función de mapeo, usando tanto reglas escritas a mano como a través de técnicas de aprendizaje automático: árboles de clasificación y regresión (CART) [Bre84], redes neuronales (NN) [McC43], aprendizaje basado en la memoria (MBL) [Sta86], máquinas de soporte de vectores (SVM) [Vap79], etc. También es posible aproximar la duración segmental usando modelos matemáticos que involucran sumas, productos, o sumas de productos de factores que influyen en la duración.

2.3.3. Modelado de la duración usando suma de productos

El modelado usando suma de productos (SoP) asume que la duración puede ser modelada usando la expresión 2.10, donde F es una función monótonamente creciente:

$$F(D(f_1, f_2, ..., f_N)) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(f_j) \tag{2.10}$$

donde

- F es desconocida pero estrictamente creciente.
- K es alguna colección de índices asociados con subconjuntos del conjunto de factores.
- I_i es una colección de índices de factores que ocurren en el i-esimo subconjunto y la función escala.
- $S_{i,j}$ es simplemente una función de mapeo entre valores discretos a valores numéricos.
- f_j representa el valor observado del factor asociado.

A través de esta expresión se pueden modelar tanto los efectos aditivos como multiplicativos de distintos factores que influyen en la variación de la duración.

Este modelo es una generalización de otros modelos. Eligiendo $K = 1, 2, ..., N$, $I_i = i$ y $F(x) = x$ conduce a varias expresiones del modelo aditivo propuesto originalmente

por Klatt [Kla76]. Alternativamente, eligiendo $K = 1$, $I_1 = 1, 2, ..., N$ y $F(x) = \log(x)$ obtenemos los modelos multiplicativos tal como se describen por Van Santen [San94]. Los modelos de suma de productos fueron aplicados al alemán (Möbius et al. [Möb96]) y al catalán (Febrer et al. [Feb98]), entre otros.

En la práctica, los métodos de SoP están relacionados con el análisis de regresión lineal múltiple tanto en el dominio lineal como el logarítmico, dependiendo de la transformación elegida. En el artículo de Van Santen [San92] se puede observar el procedimiento para el análisis de la influencia de los distintos factores, con el objeto de determinar su relevancia y comportamiento, ya sea aditivo o multiplicativo.

En el trabajo presentado por Silverman y Bellegarda [Sil99] se propone una mejora al modelo anterior a través del uso de una función sigmoide (ecuación 2.11).

$$F(x) = \left[1 + e^{-\alpha\left(\frac{x-A}{B-A} - \frac{1}{2}\right)}\right]^{-\beta} \tag{2.11}$$

donde A y B denotan las duraciones máximas y mínimas observadas en los datos de entrenamiento, y los parámetros α y β controlan la forma de la transformación. Específicamente, α controla la pendiente de la curva en el punto de inflexión, y β controla la posición del punto de inflexión en el rango de duraciones observado. Los valores de α y β se ajustan usando un algoritmo de gradiente para cada clase de fonema.

El uso de la sigmoide permite que el enfoque de SoP aproxime el 85 % de la desviación estándar con alrededor de 2,000 parámetros. Para aproximar la misma desviación usando la transformación logarítmica son necesarios más de 4,500 parámetros.

2.3.4. Modelado de la duración usando CART

Uno de las herramientas más usadas en la literatura para el modelado de la duración son los árboles de clasificación y regresión (CART, [Bre84]).

El modelo de duración del sistema de conversión texto a voz de IBM [Eid03] hace uso de un árbol de regresión (CART) para predecir la duración de los fonemas. Para cada uno de ellos, se deriva un conjunto de características del texto, tales como la identidad y características articulatorias del fonema y sus adyacentes, el número de sílabas de la palabra, la posición de la sílaba a la que pertenece el fonema y su acento léxico, distancia al final de la frase, el POS (Part-Of-Speech o categoría morfosintáctica) de la palabra a la que pertenece el fonema, etc. Estas características se usan para predecir la $\log(d)$, donde d es la duración del fonema, ya que se supone que la distribución de la duración es log-normal.

Este mismo enfoque de árboles de regresión ha sido usado por muchos otros sistemas, tales como Festival [Tay98]. En Festival coexisten dos modelos de duración que utilizan CART. El primero de ellos realiza la predicción de la duración de los segmentos en forma directa. El segundo algoritmo utiliza una adaptación de la propuesta de Campbell [Cam91], utilizando z-scores para predecir de manera indirecta la duración segmental utilizando el número de desviaciones estándar con respecto a la media.

Uno de los requisitos de este método es la disponibilidad de las medias y dispersiones para cada fonema. De esta manera, es posible predecir la duración segmental utilizando la siguiente expresión: duracion=media+(z-score . desviacion estandar).

2.3.5. Modelado de la duración usando redes neuronales

En la literatura se pueden encontrar enfoques que usan otras técnicas de aprendizaje automático, tales como redes neuronales. Para el modelado de la duración segmental en el español, Cordoba et al [Cor01] proponen el uso de redes neuronales para predecir la duración de un fonema dado un conjunto de parámetros, tales como identidad del fonema, acento léxico, posición en la frase, tipo de frase, etc.

Por otra parte, Lopez-Gonzalo [LG94] propusieron un modelado conjunto de la duración y la entonación usando un conjunto de características para definir los patrones prosódicos de la sílaba (PSP: Prosodic Syllabic Patterns): posición del acento (tres posiciones), tipo de palabra prosódica (inicial, media o final) y tipo de proposición (9 tipos). Estas son usadas para agrupar las diferentes duraciones y contornos de entonación, los cuales se representan usando dos duraciones y tres valores de f_0. En otro artículo, Lopez-Gonzalo et al. [LG96, LG97] propusieron algunas modificaciones al modelado conjunto para reducir el tamaño de la base de datos prosódica.

La predicción conjunta también ha sido propuesta por Sonntag et al. [Son97] usando redes neuronales. En su trabajo todas las redes neuronales estaban conectadas usando una estructura *feed-forward*, entrenadas usando el método estándar de *backpropagation*. Cada red contenía una capa oculta con la misma cantidad de neuronas que la capa de entrada. El número de neuronas de la capa de entrada dependía del número de parámetros elegido (entre 1 y 17). Las redes que estimaban la duración de la sílaba tenían una sola salida que representaba su duración en el rango de 60-500 ms.

2.3.6. Modelado segmental y suprasegmental

Es posible modelar la duración segmental a través del modelado de la duración suprasegmental. Campbell [Cam93] propuso que debido a que las duraciones de las sílabas se pueden predecir con un alto grado de exactitud considerando solamente un número pequeño de factores lingüísticos de alto nivel, las duraciones segmentales se pueden predecir usando un valor de elongación relativo a la duración silábica. Este proceso de acomodación de las duraciones segmentales en un marco de isocronía silábica ha sido descrito bajo la hipótesis de elasticidad por Campbell e Isard [Cam91]. Cada segmento que forma parte de la sílaba se acomoda a la duración silábica de acuerdo a su elasticidad.

La duración asignada a cada segmento dentro de la sílaba se determina mediante la siguiente fórmula:

$$\Delta = \sum_{j=1}^{n} (\mu_j + k\sigma_j) \tag{2.12}$$

donde k es una constante que se determina por un método iterativo para cada sílaba, Δ es la duración de la sílaba, n es el número de segmentos de dicha sílaba, y μ_j y σ_j son la media y la desviación estándar observadas en la base de datos para el segmento j.

Los diferentes contextos que alteran la duración silábica también fueron considerados por Campbell, tales como la variación de la velocidad del habla, el acento, y la proximidad a una frontera prosódica [Cam92a].

Un trabajo también relacionado con el modelado de la duración segmental usando unidades suprasegmentales es el realizado por Barbosa y Bailly [Bar94]. En su artículo describen la utilización de una unidad rítmica diferente a la sílaba: el *inter-perceptual center group*, unidad que puede ser detectada usando únicamente elementos acústicos [PM89]. Los autores describen un modelo para distribuir la duración del IPCG entre sus segmentos constituyentes y la incorporación dentro del proceso de la generación automática de pausas, todo ello considerando la velocidad del habla.

2.4. Junturas terminales

La división de un discurso en frases más pequeñas usando junturas terminales es uno de los temas claves relacionados con la lingüística en las tecnologías de conversión texto a voz. El principal objetivo es aumentar la inteligibilidad y mejorar la interpretación de la oración. La presencia o ausencia de una juntura terminal en una oración puede producir un cambio en su significado. Por ejemplo, en la oración [Bra03]:

"Mis tipos de emparedados favoritos son queso cremoso, jamón cocido y manteca y matambre."

El significado es diferente dependiendo de la posición de la juntura terminal:

"Mis tipos de emparedados favoritos son queso cremoso, <juntura terminal> jamón cocido y manteca <juntura terminal> y matambre." ***Significado: los tipos de emparedados son 1) queso cremoso, 2) jamón cocido y manteca, y 3) matambre.***

"Mis tipos de emparedados favoritos son queso cremoso, <juntura terminal> jamón cocido <juntura terminal> y manteca y matambre." ***Significado: los tipos de emparedados son 1) queso cremoso, 2) jamón cocido, y 3) manteca y matambre.***

Las junturas se caracterizan acústicamente a través de discontinuidades en el tono entre diferentes secciones de una oración [Bec86], pausas [O'M73, Mac76, Leh76, Kai92], y el aumento de la duración de la última sílaba antes de la juntura [Wig92].

En general, muchos autores coinciden en que básicamente hay dos niveles de junturas terminales: grupo y cláusula entonativa [Gar96]. Ambas difieren en la fuerza con que se percibe la juntura a través de los parámetros acústicos usados para indicarla.

La puntuación tiene una importante influencia en la prosodia y en la presencia de junturas terminales, ya que son marcadores del discurso usados por el escritor para indicar la manera en que el texto debe ser leído e interpretado. Por ejemplo, el signo de puntuación ":" se usa para indicar una enumeración. En tal situación, el hablante realiza una pausa, una entonación de final de oración y una reducción de la velocidad del habla para expresar al oyente que comenzará una enumeración. Algunos signos de puntuación, tales como los puntos de final de frase, indican siempre la presencia de una juntura terminal, mientras que en otros, como por ejemplo las comas, no siempre esto ocurre.

La necesidad de respirar es otra causa de las junturas terminales. El hablante inserta pausas en algunos lugares del discurso para respirar. En este caso, la ubicación se elige cuidadosamente para evitar confusiones en el oyente acerca del significado del discurso.

2.4.1. Modelado de las junturas terminales

El módulo de generación de junturas terminales es muy importante en los conversores texto a voz ya que otros módulos dependen del mismo:

- **Módulo de entonación.** La división en frases entonativas es esencial para el módulo de entonación. Las junturas se deben sintetizar con una forma particular de contorno de frecuencia fundamental para ser percibidas adecuadamente. Por ejemplo, en la oración "El hombre entró en la cueva,$<$ *junturaterminal* $>$ encendió la antorcha $<$ *junturaterminal* $>$ y observó que otro hombre lo estaba esperando.", la primera juntura tiene un contorno de entonación ascendente al final para evidenciar que continuará la oración.

- **Módulo de duración.** Como se mencionó anteriormente, este módulo predice la duración de cada fonema que será sintetizado. Las sílabas al final de frase deben presentar una duración mayor que las otras para indicar la presencia de una juntura terminal. Por lo tanto, esto repercutirá en la duración de los fonemas constituyentes de dichas sílabas.

- **Elisión.** Algunos sonidos pueden ser borrados en una oración debido a que es más fácil de pronunciar por parte del locutor. Sin embargo, la presencia de una juntura terminal puede provocar que no ocurra tal elisión.

En los conversores texto a voz la tarea del modelado de las junturas terminales consiste en decidir cuando una juntura debe ser colocada después de una palabra, utilizando la información contenida en el texto.

En general, con el objeto de realizar la estimación de la ubicación de las junturas terminales, se utiliza un conjuntos de características F más compactas que las palabras contenidas en el texto, tales como etiquetas morfosintácticas (POS), signos de puntuación, cantidad de sílabas y palabras, ubicación de las otras junturas terminales predichas, ubicación en la frase entonativa y en el grupo acentual, etc. Esto puede basarse en la probabilidad de la presencia de una juntura terminal (J) dadas esas características F: $P(J/F)$.

Es necesario destacar que la predicción de junturas terminales no es una tarea fácil debido a que también involucra un entendimiento del lenguaje natural que no poseen las computadoras y a la naturaleza arbitraria de algunas decisiones humanas. Por lo tanto, tal como ocurre con el modelado de los otros parámetros prosódicos, el conjunto de características disponibles está limitado a la capacidad de análisis del texto por parte del ordenador.

En las siguientes secciones se detallan distintos enfoques utilizados para el modelado de las junturas terminales, que involucran distintas técnicas de aprendizaje automático.

2.4.2. Modelado de las junturas terminales usando CART

Ya hemos visto que los árboles de clasificación y regresión son ampliamente usados para el modelado en los distintos componentes de un conversor texto a voz (por ejemplo, entonación y duración), y las junturas terminales son otro ejemplo de ello.

Prieto et al. [Pri96] proponen entrenar un árbol de decisión para determinar la presencia de una juntura mediante un conjunto de características contextuales de la palabra analizada, tales como: una ventana de cuatro POS, una ventana de dos palabras con información sobre acentos, el número total de palabras y sílabas de la oración, la distancia de la palabra desde el comienzo y desde el fin de la oración en palabras, sílabas, y sílabas acentuadas; distancia desde el último signo de puntuación en palabras; información sobre si la palabra está al comienzo, al final, o dentro de un sintagma nominal, su tamaño y distancia en palabras desde el comienzo del sintagma nominal. Koehn et al. [Koe00] propusieron una modificación del sistema anterior incorporando características sintácticas, y reportaron una mejora significativa. En esta misma dirección, Navas et al. [Nav02a] proponen un método basado en CART para asignar junturas terminales en euskera, usando información sintáctica y morfológica.

Estos tres métodos permiten colocar junturas terminales tomando en cuenta información local, y en el caso de Navas también se usa información acerca de la juntura predicha previamente para la decisión. La falta de esta última información podría conducir a la aparición de junturas terminales a distancias no adecuadas debido al desconocimiento de la ubicación de las junturas vecinas.

2.4.3. Modelado de las junturas terminales usando Bayes

Black et al. [Bla97] propusieron un sistema diferente basado en la regla de decisión de Bayes, el cual incorpora la ubicación de las junturas predichas previamente en la decisión. Ellos proponen maximizar la expresión

$$J(C_{1,n}) = \underset{j_{1,n}}{\operatorname{argmax}}\, P(j_{1,n}|C_{1,n})$$

donde $J(C_{1,n})$ es la secuencia de n posiciones entre palabras que pueden tener junturas terminales o no. C_i es la información de contexto de la juntura, la cual considera dos etiquetas de POS previos y la siguiente a la posición que está siendo evaluada.

$P(j_{1,n}|C_{1,n})$ se calcula como

$$P(j_{1,n}|C_{1,n}) = \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-l} \cdots j_{i-1})$$

donde $P(j_i|C_i)$ es la probabilidad que exista una juntura de acuerdo a las etiquetas adyacentes, $P(j_i)$ es la probabilidad de las etiquetas de juntura y no-juntura, y $P(j_i|j_{i-l} \cdots j_{i-1})$ es el n-grama de la probabilidad de la existencia de junturas de acuerdo a la secuencia previa de l junturas y no-junturas.

Sun et al. [Sun01] extendieron el enfoque de Black y Taylor estimando las probabilidades $P(j_i|C_i)$ usando árboles de decisión binaria. Esto permite mejorar la precisión con la que se calcula esta probabilidad, ya que puede incorporar más fuentes de información.

2.4.4. Modelado de las junturas terminales usando redes neuronales

Müller et al. [Mül00] observaron que en el modelado de las junturas terminales existe el problema de que una clase muy numerosa en los datos de entrenamiento dominará el proceso de aprendizaje.

Ellos propusieron como solución la utilización de clasificadores neuronales basados en autoasociadores, los cuales no sufren este problema, debido a que cada clase es aprendida independientemente. En consecuencia, la dispersión de los datos de entrenamiento no presenta un problema, y la capacidad de clasificación es extremadamente alta.

En el caso de Stergar et al. [Ste03], utilizaron perceptrones multicapa. Tanto el vector de entrada como el conjunto de etiquetas morfológicas fue similar al utilizado por Müller. Los autores comparan su enfoque con el de Müller, y señalan que el rendimiento es equivalente al obtenido por los clasificadores neuronales basados es autoasociadores, pero con una estructura más simple.

2.4.5. Otros algoritmos propuestos para el modelado de las junturas terminales

Varios métodos usando modelos estadísticos fueron estudiados por Sanders y Taylor [San95]. Los algoritmos utilizaban estadísticas de la aparición de junturas terminales basadas en trigramas de POS (método 1), la distancia con respecto a la última juntura predicha (método 2 y 3), y búsqueda exhaustiva de las mejores posiciones para las junturas usando una y dos fases.

Fordyce et al [For98] propusieron el uso del aprendizaje basado en reglas de transformación (TRBL: Transformational Rule-based Learning), el cual es robusto para condiciones de entrenamiento con desbalance entre las clases. Este algoritmo fue propuesto originalmente por Brill [Bri95] para el etiquetado morfosintáctico de textos. Este algoritmo produce un conjunto de reglas que transforman un texto sin etiquetar en uno con junturas terminales. Este conjunto de reglas deben ser aplicadas en forma secuencial, y se obtienen en la fase de entrenamiento minimizando el error global de clasificación.

2.5. Conclusiones

En este capítulo se ha hecho una revisión del estado de la cuestión en lo referente al modelado de distintos parámetros prosódicos relevantes para la conversión texto a voz, tales como la entonación, la duración segmental y las junturas terminales.

2.5.1. Entonación

En el caso del modelado de la entonación se han abordado distintos enfoques que abarcan un amplio espectro, tales como el modelado fonológico, el modelado perceptual y el modelado por estilización acústica. De los distintos enfoques estudiados se hizo especial énfasis en el modelado por estilización acústica tanto superposicional como no superposicional, ya que las aportaciones al modelado de la entonación de esta tesis apuntan a dichos modelos.

Sin embargo, es necesario destacar que sería deseable una unificación de los distintos enfoques para el modelado de la entonación, ya que cada uno tiene fortalezas que son complementarias a los otros modelos. El modelado fonológico tiene como fortaleza la abstracción de la función con respecto a la forma, que permite hacer un análisis de la entonación de una manera más compacta. Por otra parte, el enfoque perceptual considera solamente aquellas excursiones del contorno de entonación que son perceptibles, evitando así el intento de análisis de fenómenos no audibles. Finalmente, el enfoque de estilización acústica intenta relacionar en forma directa el contenido textual y la forma del contorno de entonación, ya sea utilizando en forma complementaria una representación abstracta (por ejemplo: ToBI) o no.

2.5.2. Duración segmental

En la sección sobre modelado de la duración segmental se han estudiado los distintos factores que influyen en su variación: segmentales, silábicos y suprasilábicos.

El modelado de la duración segmental en la conversión texto a voz ha sido abordado con un amplio espectro de enfoques de aprendizaje automático: árboles de clasificación y regresión, redes neuronales y suma de productos.

En la sección referida al modelado de la duración segmental condicionada a la duración suprasegmental, se han revisado los enfoques propuestos por Campbell y Barbosa. Campbell utilizó la sílaba como unidad suprasegmental que condiciona la duración segmental. De esta manera, una vez predicha la duración de la sílaba, esta se repartía entre los segmentos que la constituyen. El modelo propuesto por Barbosa es similar, pero haciendo uso de una unidad con mejores propiedades de isocronía (el *inter-perceptual center group*).

Este último enfoque que combina los niveles segmental y suprasegmental resultan de especial interés en la presente tesis. En el próximo capítulo, en la sección referida a las propuestas para el modelado de la duración, se expondrán algoritmos que utilizan dicho enfoque.

2.5.3. Junturas terminales

En la sección sobre junturas terminales se ha revisado el estado de la cuestión sobre su modelado para la conversión texto a voz, con propuestas que abarcan un amplio rango de técnicas de aprendizaje automático.

Las propuestas que resultan de mayor interés son aquellas que combinan en la predicción de las junturas terminales tanto la información contextual como las decisiones sobre la ubicación de junturas terminales anteriores y posteriores, como es el caso del modelo de Black et al. [Bla97].

La utilización de información local para la toma de decisiones podría conducir a la predicción de junturas terminales muy próximas entre sí, con la consiguiente merma en la naturalidad de la voz sintética.

Los modelos que combinan ambos tipos de información serán la base para los algoritmos estudiados en el siguiente capítulo, que utilizan árboles de decisión binaria, modelos de lenguaje, y transductores de estados finitos.

Capítulo 3

Aportaciones en el modelado prosódico

En esta tesis se exploran una serie de algoritmos para el modelado de la prosodia con el objeto de ser utilizados tanto para la conversión texto a voz, como para la generación de voz en el marco de la traducción voz a voz.

En la Sección 3.1 se estudiará el modelado de la entonación usando un nuevo enfoque para el entrenamiento: JEMA (Joint Extraction and Modelling Approach). Allí se estudiarán las debilidades de algunos métodos de entrenamiento propuestos en la literatura, y se detallará la metodología seguida para evitarlas. Los algoritmos propuestos para el modelado de la duración se detallan en la Sección 3.2. En esta sección se extrapolará el enfoque de entrenamiento JEMA aplicado a la entonación para mejorar el modelado de la duración usando dos niveles: segmental y suprasegmental. Finalmente, en la Sección 3.3 se explicarán los distintos enfoques para el modelado de las junturas terminales.

Las aportaciones realizadas en el modelado de la entonación, duración segmental y junturas terminales serán evaluados bajo las mismas condiciones experimentales para explorar sus fortalezas y debilidades en el Capítulo 4.

3.1. Modelado de la entonación

En la literatura se pueden encontrar muchos modelos de entonación entrenados usando dos pasos: parametrización de contornos de f_0 y entrenamiento del modelo (Sección 2.2.5). Es común la existencia de un paso previo que implica un suavizado de los contornos originales, con el objeto de eliminar ruido y microprosodia, e interpolar las regiones no sonoras para obtener un contorno de entonación continuo. El esquema de entrenamiento se puede representar en forma resumida en la Figura 3.1.

El suavizado, la interpolación de regiones no sonoras y la parametrización implican una serie de suposiciones que pueden generar problemas en el modelo de entonación, tal como se observará en la siguiente sección. Luego, se abordará el estudio del algoritmo propuesto (JEMA) en la Sección 3.1.2 para evitar los problemas inherentes de tales suposiciones.

Figura 3.1: Entrenamiento en dos pasos independientes: parametrización y entrenamiento del modelo.

3.1.1. Problemas de la parametrización

En esta sección vamos a discutir la influencia del suavizado y la interpolación en la extracción de parámetros. Para ello utilizaremos un contorno para los ejemplos que corresponde a la frase "Anda convulso el olimpo de las finanzas", que se muestra en la Figura 3.2. Para la ejemplificación usaremos el modelo de entonación de Fujisaki con las consideraciones lingüísticas propuestas por Möbius y otros [Möb95, Agü04b] que limitan la cantidad de comandos de frase y acento. Dicha limitación consiste en un comando de acento para cada grupo acentual y un comando de frase para cada grupo entonativo.

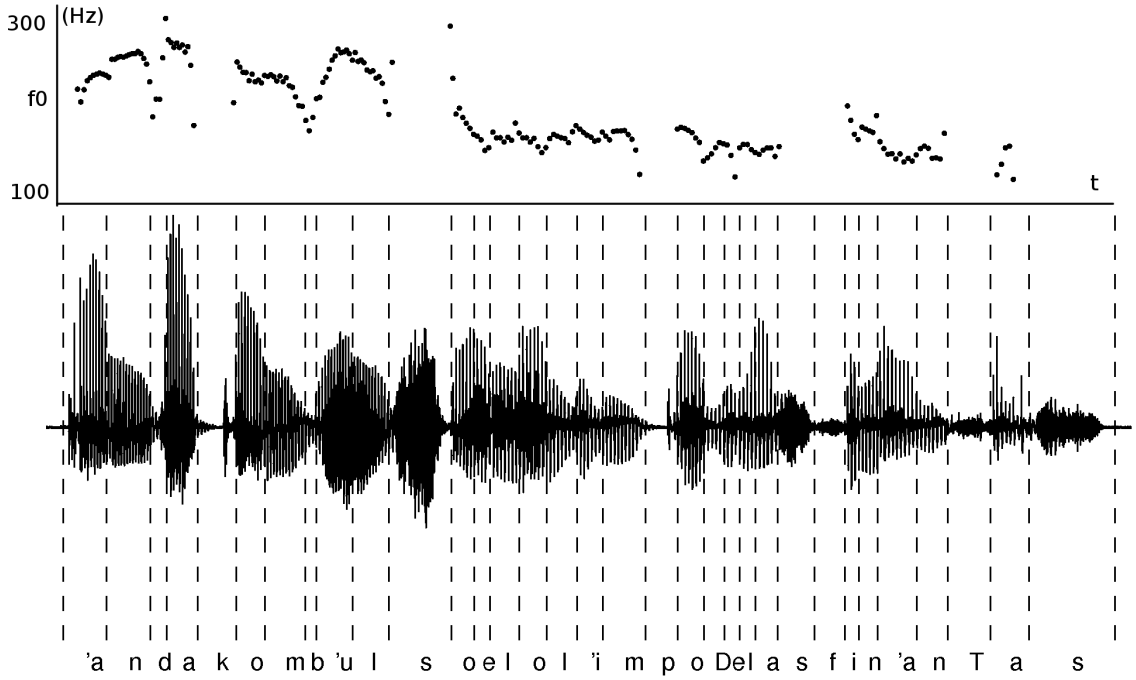
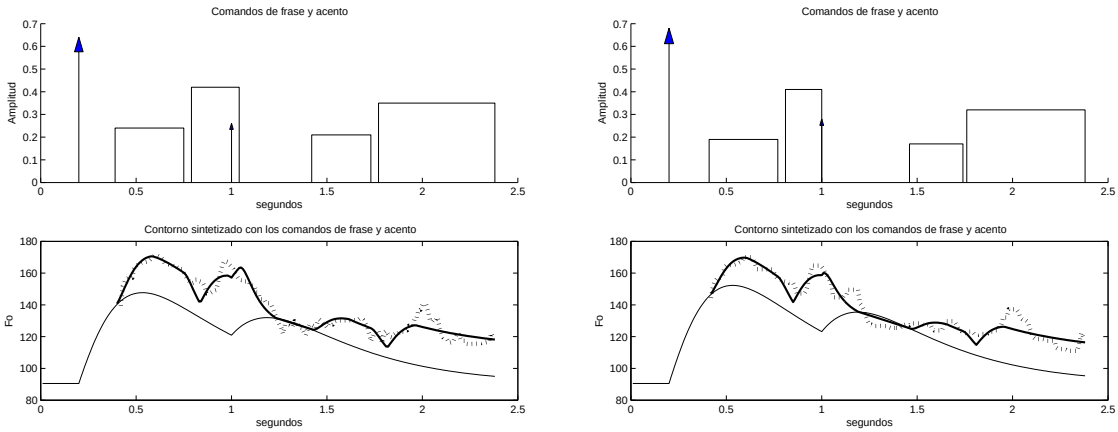


Figura 3.2: Contorno de ejemplo que corresponde a la frase "Anda convulso el olimpo de las finanzas."

La extracción del contorno de frecuencia fundamental de una señal de voz es una tarea propensa a errores, tales como *pitch halving* y *pitch doubling*. Para eliminarlos se usan técnicas de filtrado y suavizado, que pueden llegar a introducir un nuevo ruido procedente de la manipulación de la señal. Una consecuencia directa de estas operaciones es que contornos que deberían tener la misma forma no son iguales después del filtrado. Como resultado de esto, contornos iguales podrían llegar a tener parámetros diferentes debido al sesgo introducido por la utilización de estas técnicas.

En la parte inferior de la Figura 3.3 se muestra el mismo contorno de la Figura 3.2, siendo ahora continuo debido a la interpolación. Dicho contorno se presenta tanto sin suavizado (izquierda) como con suavizado (derecha) en lineas punteadas. En trazado fino se puede observar la componente de frase del modelo de Fujisaki, mientras que el contorno resultante de la suma de la componente de frase y acento se ha dibujado con trazado grueso. En la parte superior se pueden ver los comandos de frase (deltas) y los comandos de acento (pulsos).



Sin suavizado Con suavizado

Figura 3.3: Inconsistencia debido al suavizado

En el contorno de frecuencia fundamental sin suavizado existen varios errores de medición, tal como ocurre alrededor de los 600ms debido a una transición de fonema sonoro a fonema sordo. Si se observa el mismo contorno suavizado con la aplicación de un filtro de mediana de 9 muestras, este presenta una apariencia más suave debido a la acción del filtrado. El error de medición alrededor de los 600ms se ha reducido, como así también la microprosodia en el intervalo que va desde 1.3 hasta 1.6 segundos.

Los comandos de frase y acento estimados para cada caso presentan grandes diferencias en algunos de ellos, tanto en los instantes de tiempo como en sus amplitudes. Por ejemplo, la amplitud del primer y tercer comando de acento es diferente dependiendo del suavizado. Lo mismo ocurre en el segundo comando de acento, donde el instante T_2 varía dependiendo de la realización del suavizado.

En consecuencia, con este simple ejemplo se puede observar el sesgo introducido en la extracción de los parámetros debido al suavizado. Dos contornos con ligeras diferencias en su forma poseen parámetros diferentes debido a este preprocesamiento.

Otra de las suposiciones es la posibilidad de obtener continuidad en el contorno de frecuencia fundamental sin sesgar la extracción de parámetros. Algunos algoritmos de extracción necesitan contornos continuos para poder realizar la parametrización. Las técnicas de interpolación proporcionan valores de frecuencia fundamental a regiones no sonoras del habla que de otra manera no podrían tenerla. Una consecuencia de este procedimiento es que los contornos resultantes tendrán alteraciones en su forma que puede llegar a sesgar la extracción de parámetros.

Un ejemplo de ello se puede observar en la Figura 3.4, con dos contornos de la misma oración (Figura 3.2) con ligeras diferencias debido a diferentes algoritmos de estimación del contorno de frecuencia fundamental en las regiones de fonemas sordos. La interpolación de los segmentos sordos es diferente en cada caso, y se muestra en lineas punteadas.

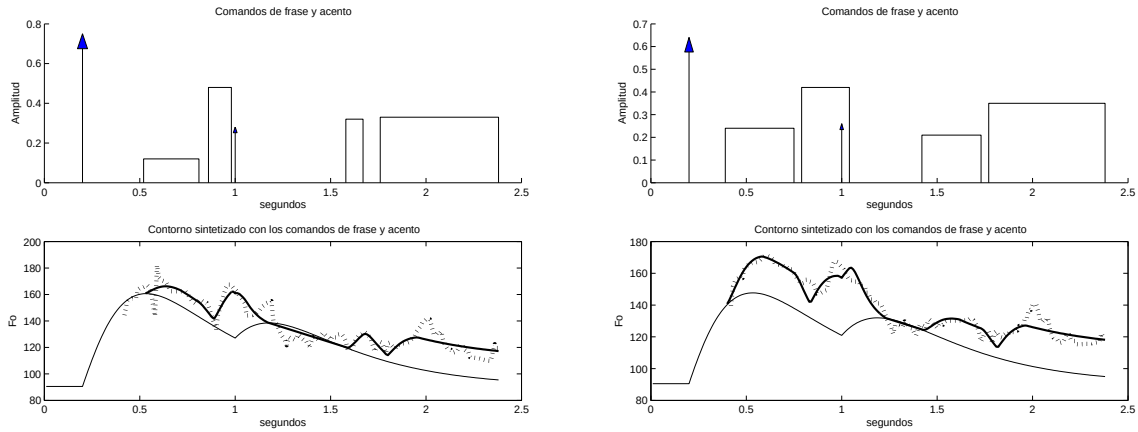


Figura 3.4: Dos ejemplos de inconsistencia en la extracción de parámetros debido a requisitos de continuidad

Cuando se realiza la estimación de los parámetros del modelo de Fujisaki, obtenemos diferentes resultados para ambos contornos debido al efecto de la información faltante en los segmentos sordos, el algoritmo de estimación de la frecuencia fundamental y el tipo de interpolación utilizado.

Es esas figuras se puede observar nuevamente que los parámetros estimados son muy diferentes, tanto en las amplitudes como en los instantes de tiempo. Esta diferencia en los parámetros extraídos debido a la interpolación y al suavizado observado en estos dos ejemplos introduce ruido en el aprendizaje automático de las reglas que relacionan las características lingüísticas y los parámetros.

Finalmente, otra importante suposición es considerar que la parametrización del contorno de frecuencia fundamental puede considerarse única. Algunos modelos de entonación no pueden asegurar esto, como es el caso del modelo de entonación de Fujisaki. En muchos de ellos existen múltiples conjuntos de parámetros que proporcionan un buen ajuste al contorno de frecuencia fundamental original. Esto vuelve la tarea de predicción más difícil, porque contornos similares tendrán diferentes parametrizaciones, aumentando la dispersión de los parámetros y creando inconsistencias.

Para estudiar las múltiples soluciones del modelo de Fujisaki se realizó un experimento que consiste en aproximar el contorno de la frase de la Figura 3.2 con dos comandos de frase y cuatro comandos de acento. Como parte del experimento se realizaron N parametrizaciones del mismo contorno variando las condiciones iniciales en la búsqueda por gradiente. De esta manera se obtiene una gran variedad de conjuntos de parámetros que aproximan con un mínimo error al contorno original.

Luego, se seleccionaron M conjuntos de parámetros de los N obtenidos fijando una distancia máxima de $0{,}001\log(Hz)$ con respecto al error medio obtenido de $0{,}083\log(Hz)$.

De esta manera se obtuvieron $M = 48$ conjuntos de parámetros del total de $N = 100$.

En la Figura 3.5 se observa un ejemplo de la dispersión de los valores de las amplitudes y los instantes de tiempo del modelo de Fujisaki para aproximar el mismo contorno de frecuencia fundamental con el mismo error $(0{,}083 \pm 0{,}001 \log(Hz))$. Las amplitudes y los instantes de tiempo corresponden al primer comando de frase y al primer comando de acento.

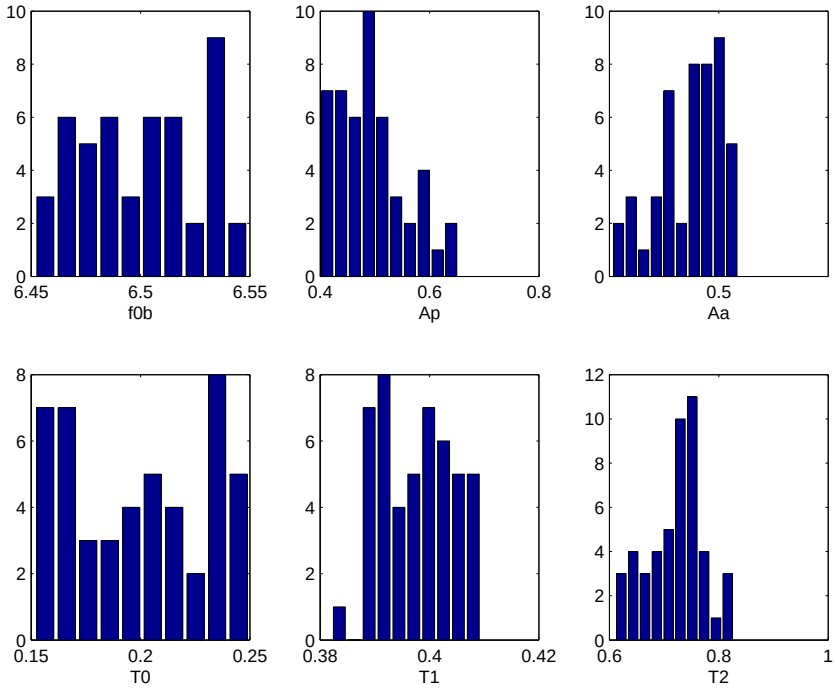


Figura 3.5: Inconsistencia originada por el tipo de parametrización

La magnitud de las diferencias, que en algunos casos puede llegar al 50 %, como es el caso de la amplitud de los comandos de frase A_p en este ejemplo, o incluso más, como ocurre con la amplitud de los comandos de acento A_a, permite darnos cuenta de la magnitud de la dispersión e inconsistencia que pueden llegar a tener los parámetros extraídos utilizando este enfoque de parametrización, introduciendo ruido en el aprendizaje automático de las reglas que relacionan las características lingüísticas y los parámetros.

3.1.2. El enfoque de parametrización y entrenamiento conjuntos (JEMA).

En esta tesis se propone un enfoque para el entrenamiento de modelos de entonación con el fin de evitar estas limitaciones. El mismo consiste en la combinación de los pasos de parametrización y entrenamiento dentro de un bucle (Figura 3.6). De esta manera, a través de sucesivas iteraciones, se obtiene una mejora en la calidad tanto de los parámetros como del modelo.

En este enfoque se hará uso de CART debido a que posee características útiles para

su uso en la conversión texto a voz, tales como la capacidad de utilizar características tanto ordinales como no ordinales, y la posibilidad para el desarrollador de interpretar fácilmente los árboles resultantes, con el objeto de introducir nuevas mejoras.

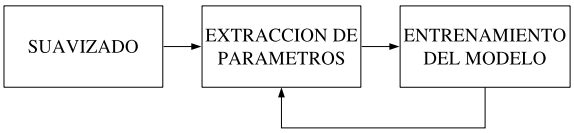
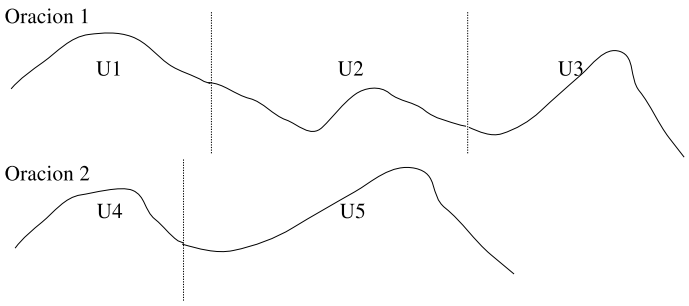


Figura 3.6: Combinación de los pasos de modelado: entrenamiendo y parametrización conjuntos.

Para explicar este enfoque se hará uso de un ejemplo. En la Figura 3.7 se observan los datos de entrenamiento consistentes en dos oraciones, con el objetivo de explicar en forma simple *JEMA* (Joint Extraction and Modelling Approach).

La primera oración tiene tres unidades prosódicas: U1, U2 y U3, y la segunda oración tiene dos: U4 y U5. Estas unidades prosódicas pueden ser grupos entonativos, grupos acentuales, sílabas, etc. Para el propósito de este ejemplo consideraremos que la unidad elegida es el grupo acentual.



Contorno de la unidad prosódica	F1	F2	F3	F4
	...	...	...	...
	...	...	...	...
	...	...	...	...
	...	...	...	...
	...	...	...	...

Figura 3.7: Ejemplo de datos de entrenamiento consistentes en dos oraciones. Las unidades prosódicas están numeradas del 1 al 5.

La base de datos de entrenamiento consiste en todas las unidades prosódicas y sus correspondientes características lingüísticas. La tabla de la Figura 3.7 muestra la base de datos de entrenamiento. Cada fila corresponde a una unidad prosódica diferente (en este ejemplo existen cinco unidades prosódicas). La primera columna es el contorno original (sin interpolación de regiones no sonoras) de la misma, y las siguientes columnas contienen las características extraídas del texto que se utilizarán para modelar los contornos.

Al comienzo, todas las unidades prosódicas en la base de datos de entrenamiento se consideran que pertenecen a la misma clase (clase 0). De esta manera, todas la unidades prosódicas se aproximan con el mismo contorno de clase (parametrización inicial). La clase 0 se representa mediante un contorno que es aquel que minimiza el error de aproximación sobre todos los contornos de entonación de las unidades prosódicas de los datos de entrenamiento usando una optimización global. La Figura 3.8 muestra la aproximación para las dos oraciones, donde se puede observar la elongación de los contornos para ajustarse a las duraciones de los diferentes grupos acentuales.

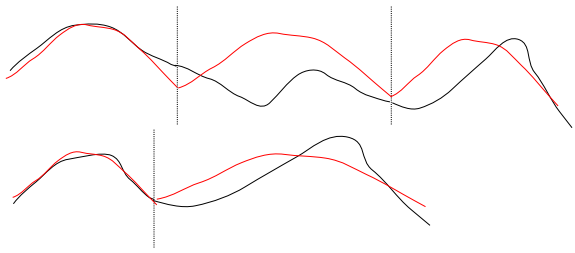


Figura 3.8: JEMA-Inicialización: Aproximación usando el contorno de la clase 0.

En el enfoque tradicional, los contornos se parametrizan uno a uno y el contorno representativo de la clase se deriva de los parámetros de cada unidad. Ya que los modelos de entonación no dependen del contenido fonético, resulta necesario que la curva resultante de la parametrización pueda ser aplicada a una unidad de entonación que tenga fonemas sonoros donde antes había sordos. Esto origina la necesidad de interpolar, tal como se mencionó en la sección anterior, para obtener un contorno de frecuencia fundamental continuo, completando la información faltante en las regiones no sonoras.

Sin embargo, si se plantea una optimización global no resultará necesario interpolar, ya que la información faltante en algunos segmentos sordos de un contorno C podrá ser estimada de aquellos contornos pertenecientes a la misma clase que poseerán segmentos sonoros en el lugar de los segmentos sordos de C. Un ejemplo de ello se observa en la Figura 3.9, donde los dos grupos acentuales son iniciales, con la misma cantidad de sílabas y el acento ubicado en la misma sílaba, pero con diferente secuencia de fonemas sonoros y sordos.

La partición de los datos de entrenamiento en subclases (fase de entrenamiento del modelo) a través de preguntas sobre las características de las unidades prosódicas permite ir reduciendo el error de aproximación (fase de extracción de parámetros). Cada posible pregunta permite definir dos clases, y al igual que con la clase 0, se busca mediante la optimización global el contorno que representa mejor cada una de ellas. Finalmente, se elige la pregunta que minimiza el error de aproximación entre los elementos de la clase y estos contornos representativos.

En la Figura 3.10 se observa la mejor partición después de intentar todas las posibles preguntas sobre las características, basada en una hipotética pregunta sobre la característica F1. La aproximación con dos clases se observa en la Figura 3.11. La partición permite aproximar bien cuatro de las cinco unidades prosódicas de la base de datos, mientras que la unidad prosódica 2 todavía no es modelada apropiadamente.

Como es habitual con las regresiones mediante árboles de decisión, este proceso de particionado continúa hasta que la condición de parada es alcanzada. El incremento del número de clases puede provocar un sobreajuste sobre los datos de entrenamiento que reduciría la generalización del modelo. Por ello es necesario usar técnicas para evitar estos efectos. Normalmente se exige un número mínimo de contornos en cada clase, y una mejora mínima medida en RMSE o correlación.

Por tanto, el proceso de entrenamiento del modelo de entonación usando JEMA consta de los siguientes pasos:

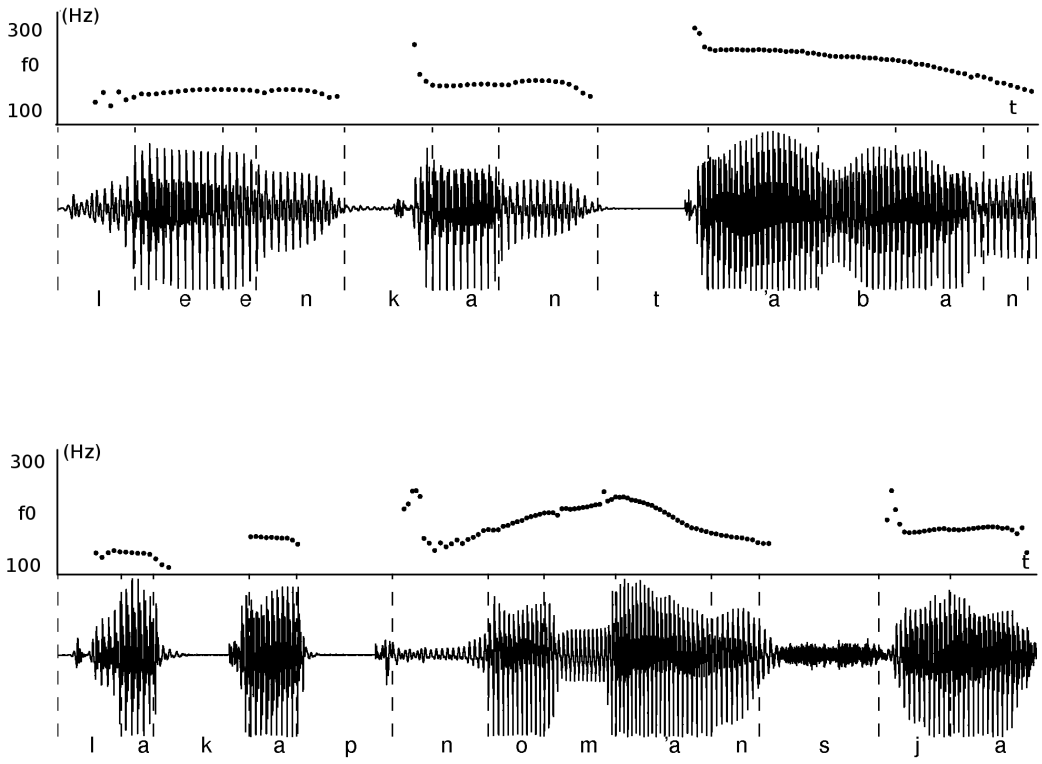


Figura 3.9: Ejemplo de complementariedad entre contornos.

Contorno de la unidad prosódica	F1	F2	F3	F4	Clase
	B	...	...	...	1
	M	...	...	...	2
	E	...	...	...	2
	B	...	...	...	1
	E	...	...	...	2

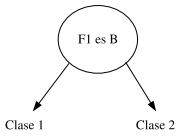


Figura 3.10: JEMA-Partición: Mejor partición en la primera iteración.

- **Inicialización.** Inicialmente solamente existe una clase, debido a que el árbol solamente tiene el nodo raíz. De esta manera, todas las unidades prosódicas serán representadas por el mismo conjunto de parámetros. Estos últimos se obtienen usando un algoritmo de optimización global sobre todos los datos de entrenamiento.

- **Partición.** Las características lingüísticas son usadas para plantear en el árbol preguntas candidatas con el objeto de dividir los datos de entrenamiento en subclases. Cada pregunta divide los datos de entrenamiento en dos nuevas clases que reemplazan a la clase particionada.

- **Optimización.** Para cada posible pregunta (partición) se usa un algoritmo de optimización global para encontrar los nuevos parámetros óptimos, tanto de las clases

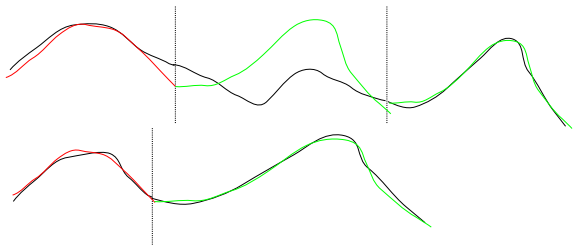


Figura 3.11: JEMA-Optimización: Aproximación con dos clases en la primera iteración.

nuevas como de las ya existentes, en caso de que sea necesario dependiendo de la parametrización.

- **Evaluación de la partición.** La nueva parametrización se utiliza para medir la mejora que introduce la partición mediante medidas objetivas, como el error cuadrático medio o el índice de correlación. Aquella partición que produzca la mayor mejora será la elegida para actualizar el árbol y pasar a la siguiente iteración.

- **Condición de finalización.** El incremento del numero de clases puede provocar un sobreajuste sobre los datos de entrenamiento que reduciría la generalización del modelo. Por ello es necesario establecer condiciones de finalización para reducir estos efectos. En este sentido, solamente se realizará otra iteración en caso de que haya clases asociadas a las hojas del árbol con un número de elementos mayor que el límite mínimo preestablecido; o bien la mejora sea menor que un umbral prefijado.

Este enfoque puede ser aplicado a varios modelos de entonación paramétricos debido a que es una técnica general para el entrenamiento de modelos de entonación. En general, su costo computacional es más elevado, ya que cada pregunta candidata exige parametrizar de nuevo los contornos de la clase. En algunos modelos esto puede resultar muy costoso, como es el caso del modelo de entonación de Fujisaki, debido al carácter no lineal del sistema de ecuaciones usado para la optimización. No obstante, con los ordenadores actuales no es un problema en la práctica.

El uso de una optimización global combinada dentro del bucle de generación del modelo de entonación evita suposiciones sobre la continuidad del contorno de frecuencia fundamental y la extracción oración por oración de los parámetros, que podría producir inconsistencias. Aquellos contornos que pertenecen a una clase poseen diferente información faltante en el contorno de frecuencia fundamental debido a la variación de la ubicación de los segmentos sordos. De este modo, para cada contorno se puede utilizar la información complementaria de los otros contornos de la clase para obtener los parámetros sin la necesidad de interpolar ni suavizar. Además, el uso de la optimización global aumenta la consistencia de los parámetros, principalmente en aquellos modelos con múltiples posibles parametrizaciones, como es el caso del modelo de Fujisaki.

En esta tesis se analiza este enfoque de entrenamiento en dos modelos de entonación: Bézier (Sección 3.1.3) y Fujisaki (Sección 3.1.4). También se ha aplicado esta técnica en el modelo de entonación Tilt tanto para el español como para el esloveno [Roj05].

3.1.3. Modelado de la entonación basado en curvas de Bézier

En este apartado se estudia la aplicación de la metodología JEMA al modelo de entonación propuesto por Escudero [Esc02b]. Dicho modelo utiliza curvas de Bezier para aproximar cada unidad de entonación, que en este caso son los grupos acentuales.

Escudero exploró varias enfoques de clasificación de los grupos acentuales según varios autores: Lopez [Lóp93], Garrido [Gar96] , Alcoba [Alc98] y Vallejo [Val98]. Cada uno de ellos consideró varios factores prosódicos que permiten determinar la forma de los grupos acentuales.

El primer paso del modelado propuesto por Escudero implica la extracción de los coeficientes de Bézier que mejor aproximan a cada grupo acentual de los datos de entrenamiento, teniendo en cuenta ciertos requisitos de continuidad. De esta manera se obtienen contornos suavizados sin transiciones bruscas entre grupos acentuales.

Luego se extraen las características lingüísticas (F) correspondientes a cada grupo acentual para clasificarlos según las propuestas de los diferentes autores estudiados. En algunos casos esta información no está disponible en un conversor texto a voz, como es el caso de los nueve tipos de grupos de entonación propuestos por Lopez. En estos casos Escudero ha realizado simplificaciones siguiendo diversos criterios detallados en su tesis.

Finalmente, el modelo de entonación utiliza un módulo de simulación y otro de generación para sintetizar el contorno de frecuencia fundamental de un grupo acentual dadas sus características. Entre las técnicas de simulación utilizadas se encuentran la simulación de distribución normal monovariable, la simulación de distribución normal multivariable, la simulación multivariable basada en datos y la simulación multivariable de aceptación/rechazo.

Debido a que en esta tesis utilizamos un árbol de decisión para representar un modelo de entonación, usaremos esta herramienta de aprendizaje automático para predecir los coeficientes de Bézier dadas las características (F). De esta manera, todos los algoritmos en estudio serán comparables utilizando la misma representación del modelo. Llamaremos a este método de estimación Bezier-SEMA (Bézier-Separate Extraction and Modelling Approach).

Bézier-SEMA presenta todas las limitaciones explicadas en la Sección 3.1.1: necesidad de interpolación de regiones no sonoras y extracción de los parámetros oración por oración.

En esta tesis proponemos estimar el modelo de entonación de Escudero mediante la técnica JEMA. Tal como hemos explicado en la sección anterior, esperamos que esta técnica mejore las prestaciones del modelo. Llamaremos a esta nueva estimación Bezier-no superposicional-JEMA.

Por otra parte, la técnica JEMA permite extender el modelo de Escudero a un modelo superposicional en el que el contorno es la suma de una componente de frase y otra de acento, ambas representadas por polinomios de Bezier. Llamaremos a este modelo Bezier-superposicional-JEMA.

En las siguientes secciones se describen los dos tipos de modelos de entonación propuestos entrenados usando el enfoque JEMA : no-superposicional y superposicional. En el

primero, la unidad prosódica es el grupo acentual, tal como ocurre en el trabajo de Escudero. Mientras tanto, en el segundo las componentes corresponden al grupo acentual y al grupo entonativo.

Modelado de la entonación basado en Bézier no-superposicional entrenado usando el enfoque JEMA

Este modelo de entonación se entrena usando JEMA, tal como se explicó en la Sección 3.1.2. Aquí detallaremos el algoritmo de optimización de los parámetros de Bézier del paso 3 de JEMA.

Dado un vector f_0, que es el vector resultante de la concatenación de todos los contornos de entonación de los datos de entrenamiento, la aproximación modela cada grupo acentual mediante un polinomio de Bezier. Se puede expresar matricialmente como

$$\hat{f}_0 = G_a a_a \tag{3.1}$$

siendo $\hat{f}_0$ la aproximación al contorno real f_0.

G_a es la matriz con los contornos de entonación g de cada orden del polinomio (0 a N, como se mostró en la Figura 2.5) y de cada clase (0 a M) de todos los contornos concatenados desde el tiempo de comienzo del primer grupo acentual (0) hasta el tiempo final del último grupo acentual de los datos de entrenamiento (T):

$$G_a = \begin{bmatrix} g^{0,0}(0) & g^{0,1}(0) & ... & g^{0,N}(0) & g^{1,0}(0) & ... & g^{1,N}(0) & ... & g^{M,N}(0) \\ g^{0,0}(1) & g^{0,1}(1) & ... & g^{0,N}(1) & g^{1,0}(1) & ... & g^{1,N}(1) & ... & g^{M,N}(1) \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ g^{0,0}(T) & g^{0,1}(T) & ... & g^{0,N}(T) & g^{1,0}(T) & ... & g^{1,N}(T) & ... & g^{M,N}(T) \end{bmatrix} \tag{3.2}$$

a_a es el vector con los coeficientes de Bézier para cada orden del polinomio (0 a N) y cada clase (0 a M), cuya solución se desea obtener:

$$a_a^T = \begin{bmatrix} a_a^{0,0} & a_a^{0,1} & ... & a_a^{0,N} & a_a^{1,0} & ... & a_a^{1,N} & ... & a_a^{M,N} \end{bmatrix} \tag{3.3}$$

Dada una unidad de entonación particular con contorno $f_0(t_1...t_2)$ asignado a la clase k, el polinomio de Bezier de orden n en el intervalo $[t1, t2]$ está definido por la ecuación 3.4.

$$g^{K,n}(\hat{t}) = \sum_{n=0}^{N} a_a^{k,n} \begin{pmatrix} N \\ n \end{pmatrix} \hat{t}^n \left(1 - \hat{t}\right)^{(N-n)} \tag{3.4}$$

El valor $\hat{t}$ es el resultado de una transformación temporal que traslada el intervalo $[t1, t2]$ en el intervalo $[0, 1]$. Además, dicha transformación fija en 0,5 el centro del núcleo de la sílaba acentuada t_n. Con el objetivo de evitar una solución complicada, se fijan las posiciones de los puntos de Bézier de manera equidistante. De esta manera la flexibilidad del modelo se reduce pero se simplifica el cálculo de los coeficientes óptimos.

Con el objeto de obtener los parámetros a_a^T óptimos, se minimiza el error de aproximación e con respecto al contorno de entonación real f_0.

$$e = f_0 - \hat{f}_0 \tag{3.5}$$

$$e^2 = e^T e = (f_0 - \hat{f}_0)^T (f_0 - \hat{f}_0) \tag{3.6}$$

La minimización se hace derivando la función de error cuadrático e^2 con respecto a los coeficientes de Bézier a_a e igualando a cero.

$$\frac{\partial e^2}{\partial a_a} = \frac{\partial}{\partial a_a}(f_0 - G_a a_a)^T (f_0 - G_a a_a) = 0 \tag{3.7}$$

Aplicando la siguiente identidad matricial:

$$(AB)^T = B^T A^T \tag{3.8}$$

obtenemos

$$\frac{\partial e^2}{\partial a_a} = \frac{\partial}{\partial a_a}(f_0^T - a_a^T G_a^T)(f_0 - G_a a_a) = 0 \tag{3.9}$$

Aplicando la propiedad distributiva:

$$\frac{\partial e^2}{\partial a_a} = \frac{\partial}{\partial a_a}(f_0^T f_0 - f_0^T G_a a_a - a_a^T G_a^T f_0 + a_a^T G_a^T G_a a_a) = 0 \tag{3.10}$$

y luego las siguientes identidades matriciales:

$$\frac{\partial x^T B x}{\partial x} = (B + B^t)x \tag{3.11}$$

$$\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a \tag{3.12}$$

se obtiene finalmente

$$\frac{\partial e^2}{\partial a_a} = \frac{\partial}{\partial a_a}(-(f_0^T G_a)^T - (G_a^T f_0) + \left((G_a^T G_a) + (G_a^T G_a)^T\right) a_a) = 0 \tag{3.13}$$

Usando la identidad $AA^T = A^T A$ y despejando obtenemos la expresión que minimiza el error e:

$$G_a^T f_0 = G_a^T G_a a_a \tag{3.14}$$

Como se mencionó anteriormente, esta expresión será usada en cada iteración y para cada posible pregunta del árbol con el fin de obtener los parámetros óptimos. El error de aproximación se calculará según la expresión del error cuadrático e^2 para medir el grado de mejora del modelo.

Una vez encontrada la pregunta que producirá la partición con el mínimo error, los coeficientes a_a resultantes de tal árbol serán el punto de partida para la mejora en la siguiente iteración.

Los coeficientes a_a obtenidos una vez cumplida la condición de parada de JEMA serán usados por el sintetizador para generar los contornos de entonación de la conversión texto a voz. Previamente los árboles determinarán a partir de las características lingüísticas la clase m que debe utilizarse en cada grupo acentual.

Una limitación del enfoque JEMA para el entrenamiento de modelos es la imposibilidad de fijar condiciones de continuidad para todos los contornos. De esta manera, aparecerán discontinuidades en los puntos de unión de las unidades prosódicas de las componentes. Esta limitación puede causar efectos no deseados, tales como discontinuidades en el medio de una palabra en caso de usar la definición de la Sección 2.2.1 para el grupo acentual. Este problema puede minimizarse aplicando un suavizado en la discontinuidad.

Modelado de la entonación basado en Bézier superposicional entrenado usando el enfoque JEMA

En esta sección se explicará el uso de JEMA para el entrenamiento de un modelo superposicional de la entonación. La utilidad del enfoque superposicional se centra en la descomposición del contorno en varias componentes, lo cual permite su estudio y modelado por separado.

Para que el modelo de entonación superposicional puede ser entrenado usando el enfoque JEMA, es necesario hacer una modificación en sus pasos. En el modelo superposicional existen dos árboles, uno para el tratamiento de los grupos entonativos, y otro para los grupos acentuales. Los árboles son generados conjuntamente, siendo evaluados en forma alternativa en cada iteración. Por ejemplo, en la primera iteración se evalúa el árbol de grupos entonativos, en la segunda el que corresponde a los grupos acentuales, y así sucesivamente.

En este algoritmo propuesto se decidió no evaluar para cada una de las M posibles particiones del árbol de grupos entonativos cada una de las N posibles particiones del árbol de grupos acentuales debido al número $M \times N$ de variantes. Con la evaluación alternativa propuesta se reduce el número de diferentes posibilidades que se deben analizar a $M + N$.

La optimización global para obtener los coeficientes de Bézier para los grupos entonativos y grupos acentuales resulta en:

$$\hat{f}_0 = G_p a_p + G_a a_a \tag{3.15}$$

siendo $\hat{f}_0$ la aproximación al contorno real, G_p los contornos de los polinomios de los grupos entonativos, a_p los coeficientes óptimos de los polinomios de los grupos entonativos,

G_a los contornos de los polinomios de los grupos acentuales y a_a los coeficientes óptimos de los polinomios de los grupos acentuales.

El objetivo es minimizar el error de aproximación e con respecto al contorno de entonación real f_0, que es el vector resultante de la concatenación de todos los contornos de entonación de los datos de entrenamiento.

La minimización se hace derivando la función de error cuadrático e^2 con respecto a los coeficientes de Bézier a_p y a_a e igualando a cero.

$$\frac{\partial e^2}{\partial a_p} = \frac{\partial}{\partial a_p}(f_0 - G_p a_p - G_a a_a)^T (f_0 - G_p a_p - G_a a_a) = 0 \tag{3.16}$$

$$\frac{\partial e^2}{\partial a_a} = \frac{\partial}{\partial a_a}(f_0 - G_p a_p - G_a a_a)^T (f_0 - G_p a_p - G_a a_a) = 0 \tag{3.17}$$

Aplicando las identidades matriciales de la sección anterior, obtenemos la ecuación matricial para calcular los coeficientes óptimos de Bézier del modelo superposicional:

$$\begin{bmatrix} G_p^T f_0 \\ G_a^T f_0 \end{bmatrix} = \begin{bmatrix} G_p^T G_p & G_p^T G_a \\ G_a^T G_p & G_a^T G_a \end{bmatrix} \begin{bmatrix} a_p \\ a_a \end{bmatrix} \tag{3.18}$$

Esta representación polinómica del contorno de frecuencia fundamental proporciona una gran flexibilidad de modelado, que dependerá del orden N elegido para los polinomios.

Finalmente, es importante destacar que debido a la ambigüedad de la formulación superposicional, es necesario utilizar JEMA para extraer adecuadamente cada componente. No es posible extraer las componentes usando SEMA sin hacer suposiciones adicionales acerca de las características de las mismas.

En la Figura 3.12 se puede ver un ejemplo de la evolución en el modelado de la entonación basado en Bézier superposicional entrenado usando el enfoque JEMA. El contorno corresponde a una frase, aunque en el entrenamiento participa todo un corpus. La frase en particular es "¿Cómo se llamaba el caballo de Calígula?", que tiene 2 grupos de entonativos "[¿Cómo se llamaba] [el caballo de Calígula?]", y cuatro grupos acentuales "(¿Cómo) (se llamaba) (el caballo) (de Calígula?)".

La primera iteración muestra el modelado de un contorno de frecuencia fundamental utilizando dos clases para el grupo entonativo y una clase para el grupo acentual. La declinación se modela correctamente, pero en cambio, el contorno de la juntura terminal en el medio y el final de la frase no son modelados apropiadamente.

El contorno de la juntura terminal en el medio de la oración es modelado apropiadamente a partir de la iteración 5, mientras que la juntura terminal de final de frase interrogativa es modelada correctamente luego de la iteración 10. Ya existen indicios de frase interrogativa después de la iteración 5, pero la excursión es pequeña y podría ocurrir que la oración no fuese percibida como una pregunta.

En este ejemplo se puede observar que en algunas iteraciones no ocurren mejoras en la aproximación del contorno real. En dichas ocasiones se ha disminuido el error de

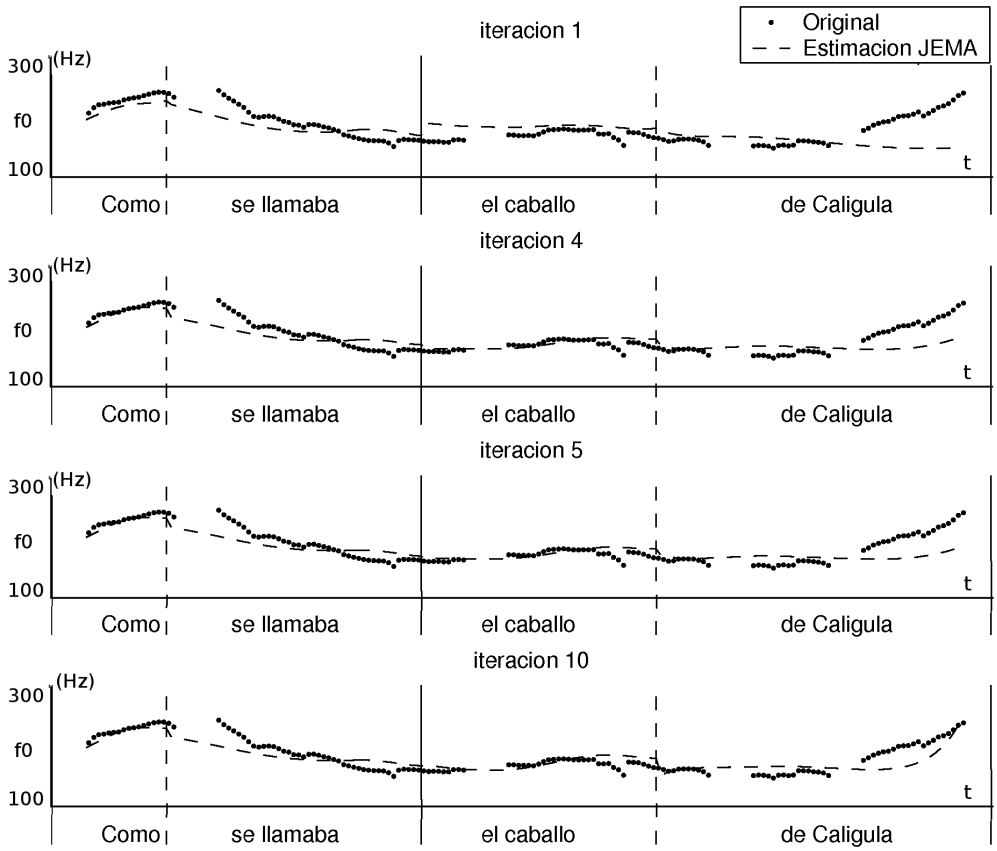


Figura 3.12: Evolución del contorno JEMA.

aproximación con respecto a otros contornos de los datos de entrenamiento, los cuales no comparten características similares con el contorno del ejemplo.

3.1.4. Modelado de la entonación usando el enfoque de Fujisaki

En esta sección se desarrolla el algoritmo de optimización global para aplicar la metodología JEMA al modelo de Fujisaki. Dado que el modelo de Fujisaki no permite encontrar una solución cerrada para hallar los comandos que mejor aproximan al contorno, tradicionalmente se han utilizado algoritmos de gradientes combinados con técnicas heurísticas [Möb95, Mix00, Nar02b, Agü04b]. Recientemente Silva et al [Sil04] presentó un método para buscar la amplitud de los comandos de acento. En este apartado proponemos una generalización para incluir la estimación de las amplitudes de los comandos de frase.

Modelado de la entonación usando el enfoque de Fujisaki aplicando JEMA

En este apartado se propone entrenar el modelo de entonación de Fujisaki usando JEMA para resolver los problemas de entrenamiento de las propuestas explicadas en la Sección 3.1.1. Con el fin de mejorar la precisión de los parámetros obtenidos, en esta sección

se incluye una formulación cerrada para el cálculo de las amplitudes. En esta formulación cerrada que se propone se asume que los instantes de tiempo son conocidos y obtenidos mediante otros procedimientos, como por ejemplo, una búsqueda basada en una cuadrícula de valores posibles, o técnicas de gradiente. En nuestro caso hemos usado el algoritmo de gradiente descendente, lo que proporciona una solución más precisa que una búsqueda en una cuadrícula.

La cantidad de comandos de acento y frase se encuentra relacionada con el número de grupos acentuales y grupos entonativos. En esta tesis se supone que para cada grupo acentual existe un comando de acento, y para cada grupo entonativo habrá un comando de frase. El tiempo T_0 de los comandos de frase es relativo al tiempo inicial del primer fonema del grupo entonativo. A su vez, el tiempo T_1 de los comandos de acento es relativo al tiempo inicial del núcleo de la sílaba acentuada del grupo acentual. El tiempo $T2$ es relativo al instante T_1, y fija la duración del comando de acento.

El bucle de optimización se muestra en la Figura 3.13. El mismo consiste en una combinación de una optimización de los valores de amplitud usando una formulación cerrada, y la actualización de los valores de los instantes de tiempo de acuerdo al gradiente. A través de sucesivas iteraciones se encuentra la solución óptima.

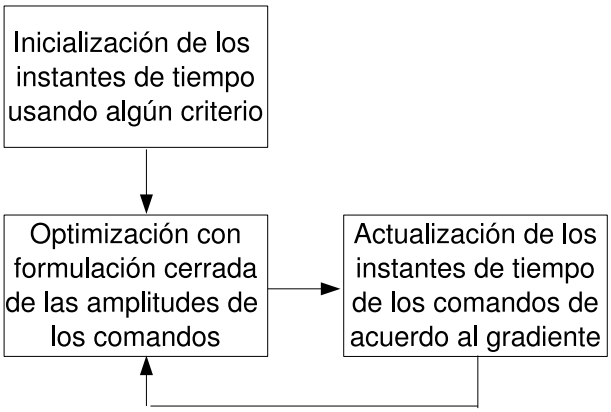


Figura 3.13: Bucle de actualización de los parámetros de los comandos de acento y frase.

Para explicar el cálculo de los valores de amplitud, haremos uso de la Ecuación 3.19:

$$\hat{f_0} = \ln f_b \, u + G_p a_p + G_a a_a \tag{3.19}$$

siendo $\hat{f_0}$ la aproximación al contorno real, f_b un escalar que representa la frecuencia base, u un vector de unos, G_p los contornos de los comandos de frase, a_p la amplitud de los comandos de frase, G_a los contornos de los comandos de acento y a_a la amplitud de los comandos de acento. Las matrices de contornos toman el valor según lo establecen las ecuaciones del modelo de Fujisaki explicadas en la Sección 2.2.5. El tamaño de los vectores a_p y a_a depende de cuantas clases posibles de comandos de frase y acento se consideren. Este número va aumentando aplicando el procedimiento JEMA explicado en la Sección 3.1.2.

El objetivo es minimizar el error de aproximación e con respecto al contorno de entonación real f_0, que es el vector resultante de la concatenación de todos los contornos de entonación de los datos de entrenamiento.

La optimización de los instantes de tiempo no se puede realizar mediante una solución cerrada debido a se encuentran en el exponente de las ecuaciones del modelo de Fujisaki. Por lo tanto, se decidió realizar la optimización mediante un algoritmo iterativo que utiliza el gradiente de la función para encontrar el valor óptimo de los parámetros. El algoritmo de gradiente descendente permite hallar el valor óptimo de los instantes de tiempo a través de sucesivas iteraciones minimizando el error e de estimación con respecto a los contornos originales. Inicialmente los instantes de tiempo de los comandos se inicializan en un valor: $T_0^0 = 0$ms, $T_1^0 = 0$ms y $T_2^0 = 50$ms. Recordemos que T_0 y T_1 están referidos al inicio del grupo entonativo y el grupo acentual respectivamente.

Luego se calcula el gradiente de la función en el instante k de acuerdo al valor de los instantes de tiempo en k ($T^k = \{T_0^k, T_1^k, T_2^k\}$), lo cual permitirá asegurar que en el instante siguiente $k+1$ los valores de los instantes de tiempo son más óptimos:

$$d^k = -\nabla e(T^k) \tag{3.20}$$

Los valores de los instantes de tiempo del instante $k+1$ se calculan en función de la dirección de gradiente descendente d^k y un parámetro α^k que debe tener un valor adecuado para asegurar la convergencia:

$$T^{k+1} = T^k + \alpha^k d^k \tag{3.21}$$

El valor de α^k se obtiene de manera tal de minimizar el error de aproximación, el cual puede ser hallado mediante una búsqueda lineal:

$$\min_{\alpha^k} e(T^k + \alpha^k d^k) \tag{3.22}$$

Formulación cerrada para el cálculo de las amplitudes de los comandos

Con el objeto de reducir el tiempo de optimización de los parámetros y aumentar la precisión de los mismos se realizaron algunas modificaciones al algoritmo propuesto por Silva [Sil04] para estimar conjuntamente los valores óptimos de la amplitud de los comandos de frase (A_p) y de acento (A_a), y de la frecuencia base ($\ln f_b$).

La optimización se hace derivando la función de error cuadrático e^2 con respecto a la amplitud de los comandos de frase a_p, la amplitud de los comandos de acento a_a y el $\ln f_b$, e igualando a cero.

$$\frac{\partial e^2}{\partial \ln f_b} = \frac{\partial}{\partial \ln f_b}(f_0 - \ln f_b u - G_p a_p - G_a a_a)^T (f_0 - \ln f_b u - G_p a_p - G_a a_a) = 0 \tag{3.23}$$

$$\frac{\partial e^2}{\partial a_p} = \frac{\partial}{\partial a_p}(f_0 - \ln f_b u - G_p a_p - G_a a_a)^T (f_0 - \ln f_b u - G_p a_p - G_a a_a) = 0 \tag{3.24}$$

$$\frac{\partial e^2}{\partial a_a} = \frac{\partial}{\partial a_a}(f_0 - \ln f_b u - G_p a_p - G_a a_a)^T (f_0 - \ln f_b u - G_p a_p - G_a a_a) = 0 \tag{3.25}$$

Aplicando las identidades matriciales de la sección anterior, obtenemos la ecuación matricial para calcular la amplitud óptima de los comandos del modelo de Fujisaki:

$$\begin{bmatrix} u^T f_0 \\ G_p^T f_0 \\ G_a^T f_0 \end{bmatrix} = \begin{bmatrix} u^T u & u^T G_p & u^T G_a \\ G_p^T u & G_p^T G_p & G_p^T G_a \\ G_a^T u & G_a^T G_p & G_a^T G_a \end{bmatrix} \begin{bmatrix} \ln f_b \\ G_p \\ G_a \end{bmatrix} \tag{3.26}$$

Este conjunto de ecuaciones permite una rápida y precisa solución de los comandos de amplitud, y evita la optimización por gradiente de tres de las seis variables en juego: las amplitudes a_a y a_p, y la frecuencia base f_b.

Es necesario aclarar que esta formulación también puede ser usada para extraer los parámetros de Fujisaki oración por oración usando el enfoque de entrenamiento SEMA. El único requisito para ello es la existencia de valores de frecuencia fundamental en todo el contorno, lo cual hace necesario el uso de la interpolación en los segmentos sordos.

3.2. Modelado de la duración

El proceso de generación de un modelo de duración es más simple que para un modelo de entonación. El objetivo del modelo es predecir la duración segmental. Por lo tanto, no es necesario seleccionar una unidad prosódica diferente al segmento.

Sin embargo, en algunos casos es útil dividir el problema de la generación de la duración de los segmentos en dos sub-problemas para tener en cuenta las características de temporización del idioma, tal como propone Campbell [Cam92a, Cam92b].

La clasificación tradicional establece que las lenguas pueden dividirse en aquellas que tienen un ritmo silábico (o isocronía silábica) y aquellas que tienen ritmo acentual (isocronía acentual). Las investigaciones que han intentado situar al español en alguno de estos extremos ponen de relieve la dificultad de su clasificación como silábica ([Alm97, Car91]) o acentual ([Man83]). Estas mismas dificultades para la clasificación pueden observarse en otros idiomas (por ejemplo, en el inglés [Bou04]).

En esta tesis proponemos la predicción de la duración suprasegmental, para luego utilizarla como información para estimar la duración de los segmentos constituyentes.

3.2.1. Predicción de la duración usando dos niveles.

El modelado de la duración segmental en base a la duración suprasegmental permitiría la utilización de la isocronía del idioma tanto para mejorar la percepción del patrón rítmico clásico del idioma, como para incrementar la precisión del modelado de la duración segmental.

En el caso del español, existen en la literatura resultados dispares. Por ejemplo, tanto Navarro Tomás [Tom22] como Borzone de Manrique y Signorini [Man83] estiman que el español debería ser catalogado como lengua de ritmo acentual, tal como ocurre con el inglés. Por otro lado, otros autores definen al español como lengua de ritmo silábico (Gili Gaya [Gay40], Delattre [Del66], Olsen [Ols72] y Carrió Eont y Ríos Mestre [Car91]).

En su trabajo de 1988, Toledo [Tol88] realizó un análisis detallado del ritmo en distintos estilos del habla por parte de hispanoparlantes sudamericanos. Como resultado de estos estudios se concluye que la duración de la sílaba se ve incrementada por tres factores: acento, posición ante la pausa y el número de los segmentos que lo constituyen.

En el mismo sentido, Pointon [Poi80] indica que no puede hablarse de ritmo en español en el sentido de producción de secuencias isócronas, tanto por lo que respecta a la sílaba como a los períodos entre acentos. Lo que existe, mas bien, es un patrón temporal condicionado por el número y tipo de segmentos en cada sílaba y la presencia o ausencia del acento.

El estudio de la duración silábica en base al número de segmentos constituyentes de los datos disponibles para los experimentos de esta tesis, revela que tanto para el locutor femenino como para el masculino la duración de la sílaba depende de dicho número. La Figura 3.14 muestra, para dos bases de datos de habla leída (locutor femenino y locutor masculino), la relación entre la duración silábica y el número de segmentos constituyentes usando los datos disponibles para los experimentos del capitulo 4. Se aprecia claramente que la duración de la sílaba crece con el número de fonemas.

Para este análisis se dispuso de la duración de 18,603 sílabas, mientras que el número de fonemas es 43,800. La silabificación se obtuvo en forma automática a través del uso de un conjunto de reglas del español. La segmentación de los fonemas en el audio fue automática usando la transcripción fonética disponible, que fue corregida manualmente, y el sistema de reconocimiento del habla de la UPC: RAMSES [Bon98]. Mediante el entrenamiento de Modelos Ocultos de Markov (HMM) de semifonemas con contexto, se determinaron las fronteras de los fonemas utilizando alineamiento forzado.

Por otra parte, tal como indican muchos autores en la literatura, es importante señalar que la duración segmental esta relacionada con la duración de la sílaba, y el número y tipo de segmentos constituyentes.

En la Tabla 3.1 se observa que en muchos casos existe una alta correlación entre la duración del fonema y la sílaba. Por ejemplo: /b/ para una sílaba con dos fonemas tiene una correlación de 0,849. Sin embargo, el mismo fonema tiene una muy baja correlación de $-0{,}046$ con la duración de una sílaba, cuando la misma tiene cuatro fonemas. Este comportamiento se corresponde con el indicado por diversos autores mencionados anteriormente.

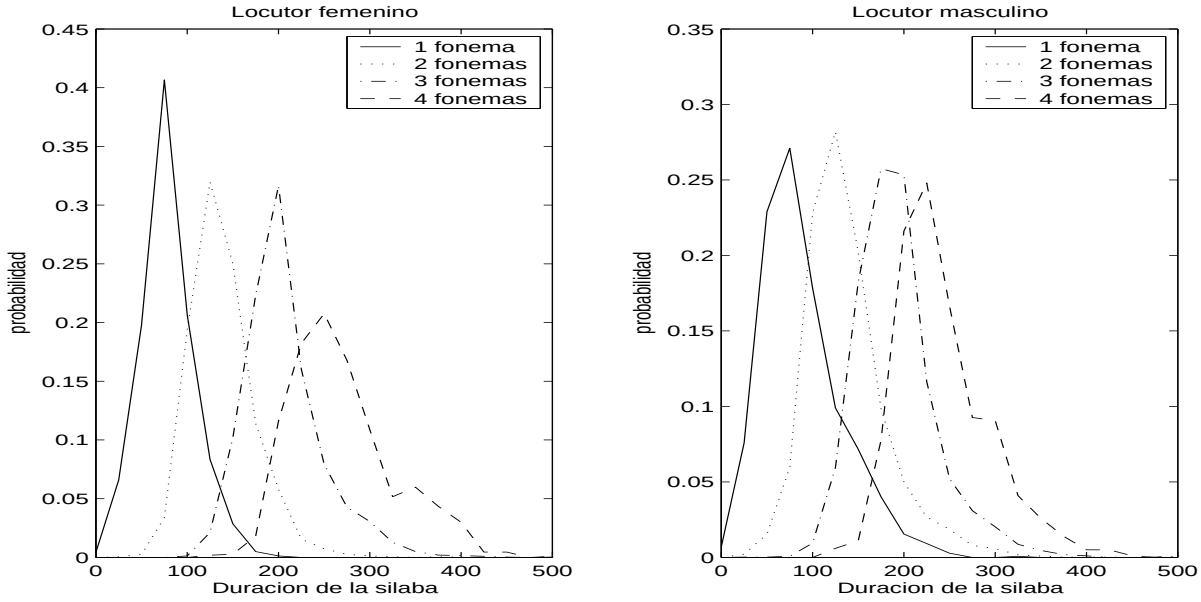


Figura 3.14: Distribución de la duración de la sílaba para diferente número de segmentos constituyentes

Fonema	**2 fon.**	**3 fon.**	**4 fon.**
a	0.692	0.670	0.560
b	0.849	0.543	-0.046
B	0.527	0.485	0.455
d	0.762	0.539	0.734
D	0.676	0.539	0.201
e	0.699	0.656	0.465
f	0.765	0.531	0.706
g	0.488	0.605	-0.049
G	0.513	0.374	0.407
i	0.699	0.618	0.813
j	0.523	0.569	0.401
J	0.631	0.737	
jj	0.663	0.439	
k	0.721	0.571	0.571
l	0.660	0.603	0.691
L	0.776	0.603	

Fonema	**2 fon.**	**3 fon.**	**4 fon.**
m	0.601	0.523	0.541
n	0.700	0.531	0.830
N	0.634	0.681	0.733
o	0.699	0.682	0.485
p	0.864	0.541	0.449
r	0.457	0.241	0.246
R	0.503	0.521	0.731
rr	0.694	0.591	0.782
s	0.734	0.778	0.769
t	0.571	0.410	0.626
T	0.771	0.576	0.369
tS	0.624	0.654	
u	0.617	0.597	0.697
w	0.683	0.633	0.637
x	0.685	0.419	0.937
z	0.908	0.776	0.699

Tabla 3.1: Análisis de la correlación entre la duración de la sílaba y la duración segmental para cada fonema discriminado por el número de segmentos constituyentes de la sílaba.

En esta tesis se propone el modelado conjunto de la duración de la sílaba (duración suprasegmental) y el fonema (duración segmental) utilizando dos enfoques.

El primero de ellos modelará la duración segmental como una fracción de la duración suprasegmental. Se estudiarán dos estimadores de los parámetros: estimación separada y estimación conjunta.

En el segundo enfoque se considerará lo observado en la Tabla 3.1 acerca de la ausencia, en algunos casos, de correlación entre la duración segmental y la suprasegmental. En este sentido se explorará un algoritmo que combina la predicción de la duración segmental como fracción de la duración suprasegmental, y también en forma absoluta.

3.2.2. Modelado de la duración segmental como una fracción de la duración suprasegmental usando estimación separada

El primer algoritmo para la predicción de la duración segmental (de los fonemas) en base a la duración suprasegmental utiliza un enfoque separado para el entrenamiento del modelo segmental y el suprasegmental.

En esta sección de la tesis se propone modelar la duración segmental como una fracción f de la duración suprasegmental, minimizando el error e con respecto a la duración real de los fonemas de los datos de entrenamiento.

En la Ecuación 3.27 se observa el cálculo del error e del modelo que debe ser minimizado ajustando los valores de las fracciones $\hat{f}_i$ y las duraciones de las sílabas $\hat{d_i^{sil}}$.

$$e^2 = \sum_{i}^{N} (d_i - \hat{f}_i . \hat{d_i^{sil}})^2 \tag{3.27}$$

La duración suprasegmental es modelada usando árboles de regresión sobre un conjunto de parámetros F, que son considerados relevantes para la estimación de la duración de la sílaba, tales como:

- Posición de la sílaba con respecto a la pausa más cercana. Por ejemplo: PREPAUSAL.
- Presencia de un acento en la sílaba. Por ejemplo: NOACENTUADA.
- Secuencia de fonemas constituyentes de la sílaba. Por ejemplo: /nes/.
- Punto de articulación de los fonemas constituyentes de la sílaba. Por ejemplo: ALVEOLAR-FRONTAL-ALVEOLAR.
- Modo de articulación de los fonemas constituyentes de la sílaba. Por ejemplo: NASAL-MEDIOCERRADA-FRICATIVA.
- Tipo de fonemas constituyentes de la sílaba (consonantes o vocales). Por ejemplo: CVC.
- Sonoridad de los fonemas constituyentes de la sílaba (sonoros o sordos). Por ejemplo: SONORO-SONORO-SORDO.
- Posición de la sílaba relativa al grupo entonativo. Por ejemplo: FINAL.

- Posición de la sílaba relativa a la palabra. Por ejemplo: FINAL.
- Número de sílabas que constituyen la palabra. Por ejemplo: 6.
- Número de fonemas que constituyen la sílaba. Por ejemplo: 3.

En cambio, la duración segmental se obtiene como una fracción de la duración silábica: $d_{fonema} = f.d_{silaba}$. En consecuencia, en lugar de modelar la duración del fonema, se modela el factor f.

El factor f también se estima usando árboles de regresión sobre un conjunto de parámetros F, que son considerados relevantes para la estimación de dicho factor, tales como:

- Características articulatorias del fonema: punto de articulación, modo de articulación, tipo de fonema y sonoridad. Por ejemplo: FRONTAL, MEDIOCERRADA, VOCAL y SONORO.
- Características articulatorias del fonema precedente. Por ejemplo: ALVEOLAR, NASAL, CONSONANTE y SONORO.
- Características articulatorias del fonema subsiguiente. Por ejemplo: ALVEOLAR, FRICATIVA, CONSONANTE y SORDO.
- Posición dentro de la sílaba: onset, núcleo o coda. Por ejemplo: NUCLEO.
- Posición de la sílaba con respecto a la pausa más cercana. Por ejemplo: PREPAUSAL.
- Presencia de un acento en la sílaba. Por ejemplo: NOACENTUADA.
- Secuencia de fonemas constituyentes de la sílaba. Por ejemplo: /nes/.
- Punto de articulación de los fonemas constituyentes de la sílaba. Por ejemplo: ALVEOLAR-FRONTAL-ALVEOLAR.
- Modo de articulación de los fonemas constituyentes de la sílaba. Por ejemplo: NASAL-MEDIOCERRADA-FRICATIVA.
- Tipo de fonemas constituyentes de la sílaba (consonantes o vocales). Por ejemplo: CVC.
- Sonoridad de los fonemas constituyentes de la sílaba (sonoros o sordos). Por ejemplo: SONORO-SONORO-SORDO.
- Número de fonemas que constituyen la sílaba. Por ejemplo: 3.

Este tipo de modelo no considera la interacción entre la estimación del valor de la sílaba y los factores de los fonemas constituyentes, debido a que los dos árboles de regresión son entrenados en forma separada. Es decir, en primer lugar se modela mediante un árbol de regresión la duración de la sílaba, para obtener el valor estimado de la misma: $\hat{d^{sil}}$. Luego,

otro árbol de regresión modela el factor $\hat{f}$, considerando la duración de la sílaba un valor conocido, obtenido mediante la estimación proporcionada por el árbol de modelado de la duración silábica.

Una consecuencia de esto es la obtención de una solución conjunta sub-óptima, la cual no solamente es originada por la característica *greedy* del entrenamiento de árboles de regresión, sino que es consecuencia de su crecimiento en forma independiente.

En las siguientes secciones se presentarán dos algoritmos para el modelado de la duración segmental como fracción de la duración suprasegmental que consideran la interacción entre los diferentes niveles en el momento del entrenamiento.

3.2.3. Modelado de la duración segmental como una fracción de la duración suprasegmental usando estimación conjunta

Uno de los principales problemas del modelado separado de la duración segmental y la suprasegmental es la tendencia al modelado de fenómenos segmentales por parte del modelo suprasegmental, o viceversa. La consecuencia inmediata de esto es una reducción de la capacidad de modelado y generalización, a causa de no considerar la interacción entre los niveles segmental y suprasegmental.

El algoritmo para la estimación conjunta de la duración suprasegmental y segmental descrito en esta sección utiliza también árboles, con el objeto de aglutinar sílabas y fonemas en grupos con una cantidad suficiente de elementos que permitan una estimación de una fracción $\hat{f}$ y una duración silábica $\hat{d}^{sil}$ adecuadas, en la Ecuación 3.27.

Como consecuencia de este agrupamiento de sílabas y fonemas, se puede reformular la Ecuación 3.27 para considerar esta nueva forma de estructurar la información. En la Ecuación 3.28 se considera la pertenencia tanto de la sílaba como del fonema a una agrupación. En el caso de la sílaba, la agrupación a la que pertenece está representada por el subíndice Cs_i, mientras que la agrupación a la que corresponde el fonema es Cf_i.

$$e^2 = \sum_{i}^{N} (d_i - \hat{f}_{Cf_i}.\hat{d}^{sil}_{Cs_i})^2 \tag{3.28}$$

Los valores óptimos para $\hat{f}_{Cf_i}$ y $\hat{d}^{sil}_{Cs_i}$ se obtienen derivando la expresión del error e^2 con respecto a dichos parámetros e igualando a cero:

$$\frac{\partial e^2}{\partial \hat{f}_{Cf_i}} = 2. \sum_{i}^{N} (d_i - \hat{f}_{Cf_i}.\hat{d}^{sil}_{Cs_i})(-\hat{d}^{sil}_{Cs_i}) = 0 \tag{3.29}$$

$$\frac{\partial e^2}{\partial \hat{d}^{sil}_{Cs_i}} = 2. \sum_{i}^{N} (d_i - \hat{f}_{Cf_i}.\hat{d}^{sil}_{Cs_i})(-\hat{f}_{Cf_i}) = 0 \tag{3.30}$$

Acomodando estas expresiones, se obtiene el conjunto de ecuaciones no lineales que

permite estimar los valores óptimos para $\hat{f}_{Cf_i}$ y $\hat{d}^{sil}_{Cs_i}$:

$$\sum_{i}^{N} d_i.\hat{d}^{sil}_{Cs_i} = \sum_{i}^{N} \hat{f}_{Cf_i}.(\hat{d}^{sil}_{Cs_i})^2 \tag{3.31}$$

$$\sum_{i}^{N} d_i.\hat{f}_{Cf_i} = \sum_{i}^{N} \hat{d}^{sil}_{Cs_i}.(\hat{f}_{Cf_i})^2 \tag{3.32}$$

Este modelo presenta ventajas en lo referente a una mejor estimación de las duraciones silábicas y los factores segmentales influyentes, debido a la optimización conjunta. Análogamente a lo explicado para el modelado de la entonación usando JEMA, en forma alternada se produce una mejora de los modelos segmentales y suprasegmentales. Es decir, en una iteración se estiman las fracciones $\hat{f}$ óptimas para las duraciones segmentales dejando constantes las duraciones suprasegmentales $\hat{d}^{sil}$. En la siguiente iteración, se recalculan las duraciones suprasegmentales $\hat{d}^{sil}$ óptimas dejando constantes las fracciones $\hat{f}$ óptimas obtenidas en la iteración anterior. Este proceso iterativo continua en forma alternada hasta que se alcanza una condición de convergencia acerca del porcentaje de mejora del error. Si la mejora del error es inferior a un umbral preestablecido, se considera que se ha convergido al valor óptimo de los parámetros $\hat{f}$ y $\hat{d}^{sil}$.

3.2.4. Modelado mixto de la duración segmental como una fracción de la duración suprasegmental y en forma absoluta usando estimación conjunta

Tal como se observó en la Tabla 3.1, no siempre es conveniente modelar la duración segmental considerando la duración suprasegmental. Por ejemplo, el fonema /b/ para una sílaba con dos fonemas tiene una correlación de 0,849. Sin embargo, el mismo fonema tiene una muy baja correlación de $-0{,}046$ con respecto a la duración de una sílaba con cuatro fonemas.

En esta sección proponemos un enfoque mixto para el modelado de la duración segmental. El mismo considera tanto su modelado en forma independiente de la duración suprasegmental, como su modelado como fracción de la duración suprasegmental.

En la siguiente ecuación se puede observar la ecuación del error de aproximación e^2 del enfoque propuesto. N representa el número de sílabas en los datos de entrenamiento, mientras que M_i corresponde al número de fonemas de la sílaba i-ésima. Es importante destacar que, a diferencia de los modelos de las secciones anteriores, en este caso el modelado se realiza a nivel de sílaba, considerando la duración segmental como parte de la duración suprasegmental.

De la misma manera que en la expresión del modelo de la sección anterior, la agrupación a la que pertenece la sílaba i-ésima está representada por el subíndice Cs_i, mientras que la agrupación a la que corresponde el fonema j-ésimo de la sílaba i-ésima es Cf_i^j.

El factor α_j^i puede tener valor 0 o 1. El mismo codifica la posibilidad de representar al fonema como una fracción de la duración de la sílaba ($\hat{f}_{Cf_i^j}\hat{d}^{sil}_{Cs_i}$), o bien en forma absoluta

($\hat{d}^{fonema}_{Cf_i^j}$).

$$e^2 = \sum_i^N \left[d_i^{sil} - \sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \right]^2 \tag{3.33}$$

Los valores óptimos para $\hat{f}_{Cf_i^j}$, $\hat{d}^{fonema}_{Cf_i^j}$ y $\hat{d}^{sil}_{Cs_i}$ se obtienen derivando la expresión del error e^2 con respecto a dichos parámetros e igualando a cero:

$$\frac{\partial e^2}{\partial \hat{f}_{Cf_i^j}} = 2 \sum_i^N \left[d_i^{sil} - \sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \right] \left(-\alpha_j^i \hat{d}^{sil}_{Cs_i} \right) = 0 \tag{3.34}$$

$$\frac{\partial e^2}{\partial \hat{d}^{fonema}_{Cf_i^j}} = 2 \sum_i^N \left[d_i^{sil} - \sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \right] \left(-(1-\alpha_j^i) \right) = 0 \tag{3.35}$$

$$\frac{\partial e^2}{\partial \hat{d}^{sil}_{Cs_i}} = 2 \sum_i^N \left[d_i^{sil} - \sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \right] \left(-\alpha_j^i \hat{f}_{Cf_i^j} \right) = 0 \tag{3.36}$$

Acomodando estas expresiones, se obtiene el conjunto de ecuaciones no lineales que permite estimar los valores óptimos para $\hat{f}_{Cf_i^j}$, $\hat{d}^{fonema}_{Cf_i^j}$ y $\hat{d}^{sil}_{Cs_i}$:

$$\sum_i^N \left[d_i^{sil} \left(\alpha_j^i \hat{d}^{sil}_{Cs_i} \right) \right] = \sum_i^N \left[\sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \left(\alpha_j^i \hat{d}^{sil}_{Cs_i} \right) \right] \tag{3.37}$$

$$\sum_i^N \left[d_i^{sil} \left(1-\alpha_j^i \right) \right] = \sum_i^N \left[\sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \left(1-\alpha_j^i \right) \right] \tag{3.38}$$

$$\sum_i^N \left[d_i^{sil} \left(\alpha_j^i \hat{f}_{Cf_i^j} \right) \right] = \sum_i^N \left[\sum_j^{M_i} \left(\alpha_j^i (\hat{f}_{Cf_i^j} \hat{d}^{sil}_{Cs_i}) + (1-\alpha_j^i) \hat{d}^{fonema}_{Cf_i^j} \right) \left(\alpha_j^i \hat{f}_{Cf_i^j} \right) \right] \tag{3.39}$$

El modelo descrito tiene una mayor flexibilidad, y por consiguiente una mejor capacidad de modelado que el descrito en la sección anterior, debido a la inclusión de dos nuevos parámetros ajustables: $\hat{d}^{fonema}$ y α. Además, comparte las ventajas de la utilización del modelado usando JEMA, que en forma alternada produce una mejora de los modelos segmentales y suprasegmentales.

Debido al caracter no lineal de las ecuaciones resultantes para la optimización, se decidió utilizar un algoritmo iterativo que utiliza el gradiente de la función para encontrar el valor óptimo de los parámetros. El algoritmo utilizado es el de gradiente descendente, ya explicado en la Sección 3.1.4 para el cálculo de los parámetros del modelo de entonación de Fujisaki.

3.3. Modelado de las junturas terminales

En la sección 2.3 del capítulo 2 se explicaron varias técnicas para el modelado de las junturas terminales orientadas al entrenamiento con datos. En esta tesis experimentaremos con tres modelos de junturas terminales con el objeto de comparar las distintas técnicas usando el mismo conjunto de datos, y de esta manera obtener una mejor noción acerca de los puntos fuertes y débiles de cada una:

- **CART.** Modelado de las junturas usando CART.
- **CART+LM.** Modelado de las junturas usando CART y un modelo de lenguaje de las junturas.
- **FST.** Modelado de las junturas usando un transductor de estados finitos.

3.3.1. Modelado de las junturas terminales usando CART.

Un enfoque básico para predecir las junturas terminales es el uso de árboles de decisión binaria. Mediante el árbol se decide si es necesario colocar una juntura terminal después de cada palabra, tal como han propuesto en la literatura varios autores [Pri96][Koe00].

Sin embargo, tal como señala Black [Bla97] en su artículo, la utilización de árboles de decisión sin tener en cuenta las decisiones previas puede llevar a la predicción de junturas terminales en palabras consecutivas cuando no es apropiado, a la falta de junturas terminales en una secuencia larga de palabras, o bien a la predicción en lugares no adecuados.

En esta tesis se propone tomar la decisión acerca de la presencia de una juntura terminal usando un conjunto de características que también consideran la ubicación de la última juntura predicha, tal como lo han propuesto Prieto et al [Pri96]. Podemos observar que cada característica utilizada se enfoca en un aspecto diferente de la modelización de las junturas terminales:

- **Ventana de POS**. Para modelar las combinaciones de POS (Part-Of-Speech, o etiquetas morfológicas) que pueda tener una juntura terminal se utiliza una ventana con tres POS a la izquierda y dos POS a la derecha. Los POS se proporcionan en forma individual y en forma agrupada, para permitir al árbol de decisión realizar una elección usando la cantidad adecuada de los mismos
- **Puntuación**. Esta característica es importante porque muchas junturas terminales están relacionadas con signos de puntuación.

- **Distancia desde la última juntura terminal.** El objetivo es evitar predecir junturas terminales en palabras consecutivas, o bien junturas terminales muy distantes.

A pesar de que el modelo es simple, los resultados son satisfactorios tal como se muestra en los resultados experimentales del capítulo 4 (Sección 4.4.2).

3.3.2. Modelado de las junturas terminales usando CART y un modelo de lenguaje.

Como se introdujo en la sección 2.4.3, Black et al. [Bla97] proponen la predicción de junturas terminales usando la regla de decisión de Bayes. Este enfoque debería aportar mejoras en las prestaciones del sistema propuesto en la sección anterior debido a que las decisiones se toman en forma óptima para toda la oración, en lugar de realizar decisiones locales.

El objetivo de este método es maximizar la probabilidad

$$J(C_{1,n}) = \underset{j_{1,n}}{\operatorname{argmax}}\, P(j_{1,n}|C_{1,n}) \tag{3.40}$$

donde $J(C_{1,n})$ es la secuencia de n decisiones sobre la presencia de junturas terminales, C_i es la información de contexto de la frontera de la palabra que es evaluada, y j_i es una etiqueta acerca de la presencia (J) o o no ($\overline{J}$) de una juntura terminal al final de la palabra.

La expresión anterior se puede escribir como

$$J(C_{1,n}) = \underset{j_{1,n}}{\operatorname{argmax}}\, \frac{P(j_{1,n}, C_{1,n})}{P(C_{1,n})} \tag{3.41}$$

donde $P(j_{1,n}, C_{1,n})$ se puede descomponer en

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} P(C_i|j_{1,i}, C_{1,i-1}) P(j_i|j_{1,i-1}, C_{1,i-1}) \tag{3.42}$$

Si se hacen algunas suposiciones como las siguientes

$$P(C_i|j_{1,i}, C_{1,i-1}) \approx P(C_i|j_i) \tag{3.43}$$

$$P(j_i|j_{1,i-1}, C_{1,i-1}) \approx P(j_i|j_{i-k,i-1}) \tag{3.44}$$

obtenemos

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} P(C_i|j_i) P(j_i|j_{i-k,i-1}) \tag{3.45}$$

Si se usa la siguiente igualdad

$$P(C_i|j_i) = \frac{P(C_i)P(j_i|C_i)}{P(j_i)} \tag{3.46}$$

se obtiene finalmente

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} \frac{P(C_i)P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-k,i-1}) \tag{3.47}$$

Como consecuencia, maximizamos la siguiente expresión

$$J(C_{1,n}) = \underset{j_{1,n}}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-k,i-1}) \tag{3.48}$$

donde $P(j_i|C_i)$ es la probabilidad de que exista una juntura terminal dado el contexto C_i, $P(j_i)$ es la probabilidad de la existencia o no de una juntura terminal, y $P(j_i|j_{i-k} \cdots j_{i-1})$ es el n-grama que describe la probabilidad de una juntura dada la secuencia de k decisiones previas sobre junturas.

La probabilidad $P(j_i|C_i)$ se ha estimado mediante CART usando la información sobre POS y puntuación, tal como propone Sun et al. [Sun01]. La información relacionada con la distancia con respecto a la última juntura terminal no se usa debido a que es considerada por el modelo de lenguaje.

3.3.3. Modelado de las junturas terminales usando transductores de estados finitos.

Este tercer algoritmo propuesto realiza una conversión de las etiquetas POS en etiquetas de junturas terminales. La ecuación de partida es la misma que la primera de la sección anterior, pero aquí se simplifican los contextos utilizando únicamente el POS. Aunque el contexto es más limitado, veremos que no será necesario realizar aproximaciones tan severas como las utilizadas en la sección anterior.

El problema se puede resolver mediante un transductor de estados finitos (FST: Finite State Transducer), cuyo lenguaje de entrada son las etiquetas POS y la salida son las etiquetas de junturas (J o $\overline{J}$). Los FST han sido usado para varias tareas, tales como transcripción fonética [Gal01] y traducción automática [Gis02]. Estas tareas son más complejas que la predicción de junturas terminales, debido a que en muchos casos hay un mapeo de varias entradas a varias salidas. Por ejemplo, varias palabras del idioma origen se traducen en otro conjunto de palabras del idioma destino: “Up to the present” → “Hasta ahora”. Además, en algunos casos la secuencia de salida tiene un orden diferente que en la entrada.

La información dada al transductor se muestra en la Tabla 3.2. Las etiquetas de la salida son $\overline{J}$ (no existe juntura terminal) o J (existe juntura terminal). La posición de las junturas terminales esta asociada al final de las palabras. Los signos de puntuación

son concatenados con las etiquetas de morfosintácticas para modelar la relación entre las junturas terminales y los signos de puntuación.

Texto	**Entrada**	**Salida**
El	DT	$\overline{J}$
rey	NN	J
está	VBZ	$\overline{J}$
tocando	VBG	$\overline{J}$
el	DT	$\overline{J}$
piano	NN,	J
mientras	IN	$\overline{J}$
la	DT	$\overline{J}$
reina	NN	$\overline{J}$
canta.	VBZ.	J

Tabla 3.2: Entradas y salidas del transductor de estados finitos. $\overline{J}$ indica que no existe juntura terminal, y J indica que existe juntura terminal.

En nuestro enfoque decidimos usar etiquetas morfosintácticas por dos razones:

- **Reducción del tamaño del espacio de entrada**. Las etiquetas morfosintácticas son usadas en lugar de las palabras. El uso de palabras provocaría la necesidad de una gran cantidad de datos de entrenamiento para obtener estimaciones confiables de las probabilidades.
- **Relaciones entre las etiquetas morfosintácticas y las junturas terminales**. Existen varios trabajos en el área que muestran que las etiquetas morfosintácticas son una importante fuente de información para decidir la ubicación de las junturas terminales [Bla97, Pri96].

Partiendo de la ecuación de la sección anterior

$$J(C_{1,n}) = \underset{j_{1,n}}{\operatorname{argmax}}\, P(j_{1,n}|C_{1,n})$$

y sustituyendo C_i por p_i,

$$J(p_{1,n}) = \underset{j_{1,n}}{\operatorname{argmax}}\, P(j_{1,n}|p_{1,n})$$

que se puede reescribir de la siguiente manera:

$$\underset{j}{\operatorname{argmax}}\, P(j|p) = \underset{j}{\operatorname{argmax}}\, \frac{P(j,p)}{P(p)} = \underset{j}{\operatorname{argmax}}\, P(j,p) \tag{3.49}$$

Aplicando la regla de bayes obtenemos la siguiente expresión:

$$P(j,p) = \prod_{i=1}^{N} P(j_i, p_i | j_{i-k}^{i-1}, p_{i-k}^{i-1}) \tag{3.50}$$

En la fase de entrenamiento, el transductor recibe una secuencia de parejas de etiquetas de POS y de presencia/ausencia de junturas:

$$(p_1, j_1)(p_2, j_2)...(p_n, j_n) \tag{3.51}$$

donde p_i es la etiqueta POS de la palabra w_i, y j_i indica la presencia o ausencia de juntura terminal (J y $\overline{J}$, respectivamente) después de la palabra w_i.

La tarea del transductor es encontrar la secuencia de etiquetas de junturas terminales que maximizan la ecuación 3.52.

$$\underset{b}{\operatorname{argmax}}\, P(j/p) = argmax_j \frac{P(j,p)}{P(p)} = argmax_j P(j,p) \tag{3.52}$$

$P(j,p)$ es la probabilidad conjunta de una secuencia de etiquetas de POS y junturas terminales. La misma se puede modelar usando n-gramas, como se muestra en la ecuación 3.53.

$$P(j,p) = \prod_{i=1}^{N} P(j_i, p_i / j_{i-k}^{i-1}, p_{i-k}^{i-1}) \tag{3.53}$$

Esta expresión se puede reescribir usando el concepto de tuplas, término usado comúnmente en traducción, y que se define formalmente como el conjunto de las frases más cortas que proporcionan una segmentación monótona de los datos bilingües. En este caso la tupla es simplemente la concatenación de la etiquetas de POS y de juntura. La expresión reescrita resulta:

$$P(t) = \prod_{i=1}^{N} P(t_i | t_{i-k}^{i-1}) \tag{3.54}$$

donde t_i es la tupla definida por el POS y la juntura i-ésima.

La probabilidad de la secuencia de tuplas se estima mediante un n-grama, y puede representarse mediante un autómata de estados finitos (FSA). Cada estado representa una historia (t_{i-k}^{i}) y los arcos contienen la probabilidad condicional de una observación dada la historia previa ($P(t_i | t_{i-k}^{i-1})$). De esta manera, la probabilidad conjunta de una secuencia de observaciones se puede obtener atravesando el autómata de estados finitos dadas las observaciones.

En esta tesis, los n-gramas son de longitud variable, tal como se propone en Bonafonte et al. [Bon96]. La idea básica es que los estados con historia ($w_{t-m}...w_t$) son candidatos

a ser combinados con los estados $(w_{t-m+1}...w_t)$, con el objetivo de obtener probabilidades más confiables para las historias más largas. Los criterios empleados para tomar esta decisión son:

- Los estados se combinan si el número de veces que la historia $(w_{t-m}...w_t)$ ha sido observada en los datos de entrenamiento es menor a un umbral.
- Los estados se combinan si la información de la distribución $p = p(w|w_{t-m}...w_t)$ es similar a la de la distribución $p = p(w|w_{t-m+1}...w_t)$.

Para convertir al autómata de estados finitos en un trasductor de estados finitos, se tiene en cuenta que la observación de la etiqueta POS p_i genera una salida j_i. Dadas las entradas p_i, existen varios caminos posibles en el FST que podrían ser atravesados con la secuencia p. Para encontrar el camino que maximiza $P(j|p)$ se usa el algoritmo de decodificación de Viterbi. Dada la secuencia de estados óptima, es posible obtener las etiquetas de juntura terminal (j_i) que corresponden al mejor camino a través del FST [Bon04].

3.3.4. Modelado de las junturas terminales usando grupos acentuales.

Marin et al. [Mar96] proponen en su artículo la utilización del concepto de grupo acentual para modelar las junturas terminales. Los autores asumen que en el idioma español no hay junturas terminales dentro de un grupo acentual. Si esta hipótesis es cierta, se puede usar en cualquiera de los otros métodos como información contextual, reduciendo el espacio de búsqueda y posibles errores.

Un grupo acentual está definido como la secuencia de palabras que pertenecen a clases morfosintácticas no acentuadas (tales como determinantes, adjetivos posesivos, preposiciones, conjunciones y pronombres no acentuados) finalizando con una palabra acentuada (sustantivo, adjetivo, pronombres acentuados, verbos y adverbios).

En la base de datos de TC-STAR (Sección 4.4.1) observamos que esta hipótesis es cierta: no existen junturas terminales después de palabras no acentuadas. Este fenómeno no resulta extraño debido a que los datos de entrenamiento se grabaron usando condiciones ideales, sin la presencia de disfluencias. Las disfluencias son de caracter aleatorio, y pueden introducir junturas terminales en cualquier posición.

En esta tesis incluiremos resultados experimentales de los métodos previos usando grupos acentuales en lugar de palabras.

3.4. Conclusiones

En este capítulo se han explicado varias propuestas para la mejora de la calidad de la entonación, duración segmental y predicción de junturas terminales.

3.4.1. Entonación

En el contexto de la generación de la entonación se ha propuesto un enfoque nuevo para el entrenamiento de los modelos: JEMA (*Joint Extraction and Modelling Approach*). La idea consiste en combinar los procesos de extracción de parámetros y generación del modelo en un ciclo de mejora continua, donde en cada iteración se refinan tanto los parámetros como el modelo.

Una de las características distintivas del enfoque es la ausencia del requisito de continuidad de los contornos de frecuencia fundamental. La extracción global de los parámetros evita la interpolación para calcular valores de frecuencia fundamental en segmentos no sonoros, y el consecuente sesgo debido a este procedimiento no estará presente.

Por otra parte, esperamos que la estimación global de los parámetros mejore la consistencia de los mismos, especialmente para modelos de entonación con una importante multiplicidad de soluciones posibles, como es el caso del modelo de Fujisaki.

3.4.2. Duración

En lo relativo al modelado de la duración segmental se ha propuesto un enfoque que combina la isocronía del idioma y su relación con la duración de los segmentos constituyentes.

A través de un estudio de los datos de entrenamiento se demuestra la dependencia entre la duración de la sílaba y el número de segmentos constituyentes. Además, tal como han explicado varios autores del área, la duración de la sílaba también depende de factores tales como la prominencia y la cercanía de una frontera prosódica. Como consecuencia de estas observaciones, se propone el modelado segmental utilizando la duración silábica, sin considerar una isocronía silábica estricta.

Los dos primeros algoritmos propuestos consideran que la duración segmental puede modelarse como una fracción de la duración silábica. En consecuencia, cada segmento variará en función de la duración suprasegmental, ajustándose todos los constituyentes a la duración predicha de la sílaba.

La observación de la correlación entre la duración de la sílaba y la duración segmental nos permite determinar que en algunas ocasiones pueden considerarse como fenómenos que no guardan una relación lineal entre ellos. Teniendo en cuenta esto, en esta tesis se propone el modelado de la duración segmental de manera condicional, considerándola como una fracción de la duración silábica, o bien en forma absoluta, independiente de la duración suprasegmental.

Los dos algoritmos propuestos utilizan una extrapolación para el modelado de la duración del enfoque JEMA utilizado para el modelado de la entonación.

3.4.3. Junturas terminales

Finalmente, la última sección del capítulo ha tratado sobre propuestas para el modelado de las junturas terminales, utilizando distintos enfoques tanto en lo referente a la formulación matemática del problema como a la unidad elegida: palabra o grupo acentual.

El primer enfoque propone una modificación a la predicción de junturas terminales usando árboles de clasificación y regresión, con la inclusión de información sobre la distancia de la última juntura terminal predicha. Dicha información proporciona información adicional con el objeto de evitar la predicción de junturas terminales muy próximas. Sin embargo, solamente se predicen de izquierda a derecha (siguiendo el sentido de lectura de la oración), sin intentar encontrar una ubicación óptima a nivel oración de las junturas.

El segundo enfoque propuesto incorpora en un modelo tanto la propuesta de Black [Bla97] como la de Sun [Sun01]. La utilización de una mayor cantidad de información contextual para determinar la probabilidad de una juntura y el uso de un modelo de lenguaje sobre la ubicación de las junturas terminales, conjuntamente con la utilización del algoritmo Viterbi, permite encontrar la ubicación óptima de acuerdo a las probabilidades estimadas.

La utilización de un transductor de estados finitos con etiquetas morfosintácticas y puntuación como entrada para la predicción de junturas terminales es una simplificación del modelo anterior en lo que respecta a la información contextual. Sin embargo, el modelo incorpora una complejidad adicional ya que incluye información contextual de longitud variable a través del uso de n-gramas.

En el siguiente capítulo se describirán tanto el marco experimental como los resultados de la aplicación de los diferentes algoritmos descritos en esta sección a los datos de entrenamiento y evaluación disponibles en las voces grabadas para el proyecto TC-STAR.

Capítulo 4

Validación experimental de las aportaciones

En este capítulo se presentan los resultados experimentales para estudiar las ventajas de los algoritmos propuestos con respecto a algunos enfoques extraídos de la literatura en lo relacionado al modelado de la entonación, la duración y las junturas terminales.

En el caso del modelado de la entonación se han hecho experimentos tanto con contornos artificiales como reales. Los primeros se utilizaron para estudiar las ventajas del método JEMA para el entrenamiento de modelos de entonación en condiciones controladas de experimentación (Sección 4.1), mientras que los segundos permitieron analizar el enfoque propuesto para la generación de contornos reales. En la Sección 4.2 se muestran los experimentos para el modelado de la entonación usando contornos reales, a través de dos algoritmos diferentes: SEMA y JEMA; y tres modelos matemáticos: S-Bézier, Bézier y Fujisaki.

Los resultados experimentales sobre el modelado de la duración segmental se presentan en la Sección 4.3, con el objeto de demostrar las ventajas de la utilización de JEMA en el entrenamiento de dicho modelos.

Finalmente, en la Sección 4.4 se muestran los experimentos realizados sobre modelado de las junturas terminales con los tres algoritmos propuestos en el capítulo anterior: árboles de clasificación y regresión, árboles de clasificación y regresión con un modelo del lenguaje, y transductores de estados finitos. Cada uno de estos algoritmos es evaluado usando dos unidades diferentes: palabras y grupos acentuales.

4.1. JEMA: una prueba de concepto

En el modelado de la entonación es importante el uso de datos reales para extraer conclusiones acerca de la calidad de los diferentes métodos de entrenamiento para la correcta generación de los contornos de frecuencia fundamental en un sistema de conversión texto a voz. Sin embargo, diversas particularidades de la prosodia pueden llevar a problemas en el proceso de comparación de diferentes métodos de estimación, y a extraer conclusiones

erróneas acerca de la precisión de los modelos.

Los errores de estimación en el contorno de frecuencia fundamental debido a la microprosodia o a limitaciones de los algoritmos de extracción introducen un ruido que puede llegar a influenciar en el rendimiento de los modelos. Es posible que la existencia de una duplicación del valor de la frecuencia fundamental (*pitch doubling*) penalice en gran medida una porción del contorno estimado debido al uso del error cuadrático medio como medida de comparación.

Otra característica que ejerce una influencia importante en el proceso de comparación de modelos son las limitaciones de la formulación matemática. Dicha formulación puede contener una capacidad de aproximación limitada para los distintos contornos de frecuencia fundamental disponibles para la comparación de diferentes técnicas de entrenamiento. Esto puede llegar a enmascarar la viabilidad de un método debido a que el nivel de error de estimación puede ser mayor que el introducido por los algoritmos de entrenamiento. Por ejemplo: en un corpus expresivo se puede observar una fluctuación con varios máximos y mínimos de la curva entonativa dentro de una sílaba, fenómeno que no puede aproximarse fácilmente si no se incluye la suficiente cantidad de parámetros en el formulación matemática.

El carácter no determinístico de los contornos de frecuencia fundamental humanos también repercute en la capacidad para comparar diferentes modelos. La habilidad de los humanos para producir diferentes contornos de frecuencia fundamental manteniendo el significado de lo expresado introduce una variabilidad importante en la tarea. Por lo tanto, el modelado o la comparación usando contornos de referencia solamente es una aproximación a la medición de la precisión de los modelos debido a la multiplicidad de contornos válidos.

Finalmente, la información insuficiente disponible para el modelado de la entonación introduce también problemas en la comparación de modelos. La información disponible para un humano es muy superior que la que puede llegar a manipular un ordenador, tales como información sintáctica, semántica, pragmática, etc. En consecuencia, los errores en la estimación están contaminados por la cantidad de características (F) no disponibles para el entrenamiento, que pueden provocar el solapamiento de muchas clases.

Debido a las razones mencionadas anteriormente es que incluimos un conjunto de experimentos con contornos artificiales para analizar el rendimiento de JEMA comparado con el enfoque SEMA. El objetivo es evaluar si en una situación ideal, con completa disponibilidad de las características necesarias para garantizar la separabilidad de las clases, JEMA aporta una mejora con respecto a SEMA. El objetivo también es observar si el modelado con este método es superior, evitando su contaminación por otros factores, como es el caso de la microprosodia y los errores de estimación de la frecuencia fundamental.

4.1.1. Datos experimentales

Los datos artificiales se generaron usando un conjunto de ocho clases arbitrarias con parámetros aleatorios. Los parámetros para cada clase se seleccionaron para obtener contornos de frecuencia fundamental en el rango de 100Hz a 200Hz, usando dos parametriza-

ciones correspondientes a modelos superposicionales: Bézier y Fujisaki. En la Figura 4.1 se puede ver un ejemplo de dichos contornos para el caso del modelo superposicional de Fujisaki.

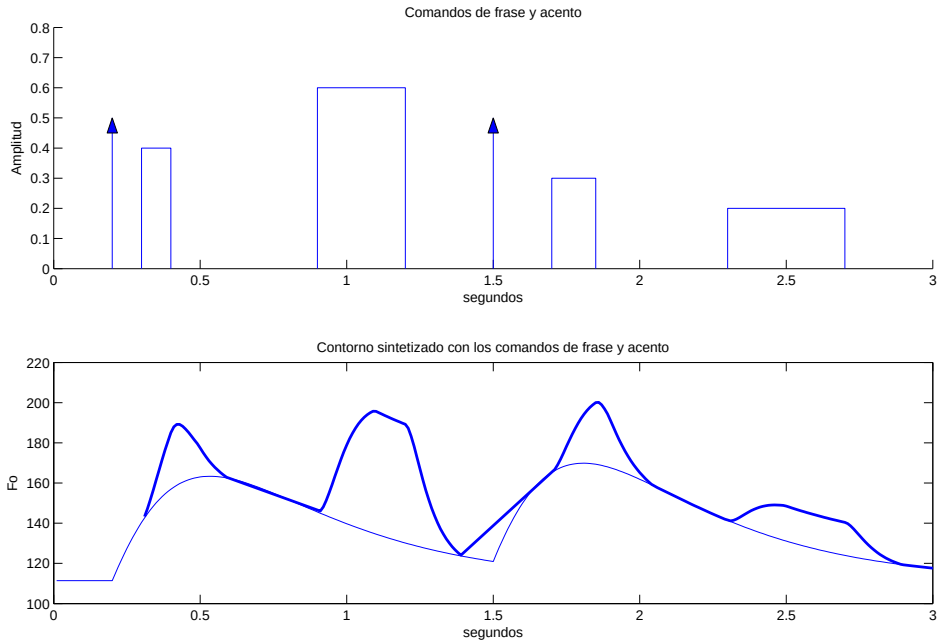


Figura 4.1: Contorno artificial correspondiente al modelo superposicional de Fujisaki.

En total se generaron cuarenta contornos de entrenamiento. La duración de cada contorno es de 2 a 3 segundos, compuestos de 4 a 8 grupos acentuales y 2 a 4 grupos entonativos. La presencia de grupos acentuales y grupos entonativos permitirá evaluar la capacidad de los algoritmos para detectar las diferentes componentes de un modelo superposicional.

Los contornos se generaron usando diferentes porcentajes de datos faltantes de manera artificial (0 % a 80 %), con el objetivo de simular la ausencia de información debido a segmentos sordos de 50ms a 100ms. Para el método SEMA se incluyó un pre-procesamiento en los segmentos sordos que consistió en una interpolación lineal y un filtro de mediana.

Además, se incluyó ruido gaussiano de media cero y dispersión σ, en el rango de 0Hz a 3Hz, con el fin de simular la presencia de ruido de estimación y de microprosodia.

Cada clase posee un conjunto de características (F) que permiten la completa separación de las clases, con el fin de evitar un solapamiento que provocaría los problemas de modelización ya explicados en la sección anterior.

4.1.2. Resultados experimentales

En la Figura 4.2 se observan los resultados experimentales usando los parámetros de Bézier, entrenados tanto con el enfoque SEMA (lineas sólidas) como el JEMA (lineas punteadas), para diferentes niveles de ruido $\sigma = 0Hz$ (diamante), $\sigma = 1Hz$ (estrella), $\sigma = 2Hz$ (cuadrado) y $\sigma = 3Hz$ (equis). El eje horizontal de la gráfica representa los diferentes niveles de datos faltantes.

Los modelos entrenados usando el enfoque JEMA tienen el mismo RMSE para cualquier porcentaje de datos faltantes. El error solamente se ve incrementado debido al ruido gausiano introducido. El mejor rendimiento de los modelos entrenados con JEMA es una consecuencia directa de la consistencia introducida en la parametrización debido al uso de una optimización global. JEMA evita los sesgos producidos por la falta de datos en los segmentos sordos.

Sin embargo, los modelos entrenados usando SEMA sufren un fuerte impacto en su rendimiento con el incremento de los datos faltantes, tal como se observa para niveles de 30 % en adelante.

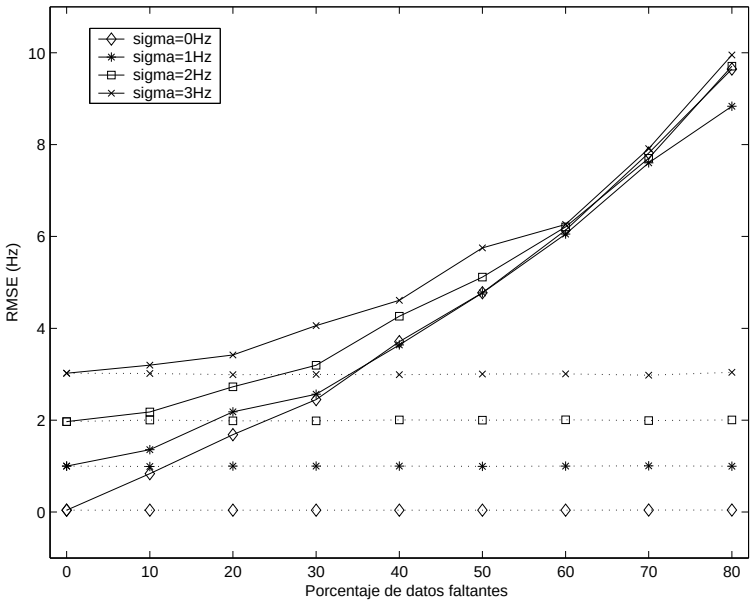


Figura 4.2: RMSE obtenido usando parametrización de Bezier para diferentes condiciones de ruido e información faltante en los datos de entrenamiento.

En la Figura 4.3 se muestran resultados experimentales usando parámetros de Fujisaki. En dicha figura se puede observar el mismo comportamiento de JEMA vs SEMA para diferentes porcentajes de datos faltantes debido a segmentos sordos y diferentes niveles de ruido. El pequeño aumento debido al ruido para JEMA se debe a la sensibilidad del modelo a los instantes de tiempo de los comandos de frase y acento: T_0, T_1 y T_2. Una pequeña diferencia puede llegar a introducir un error que dependerá de las constantes α y β elegidas. Este efecto es menos significante para mayores niveles de σ debido a la influencia más fuerte del ruido gausiano en los contornos artificiales. Se observa que JEMA supera

en rendimiento a SEMA incluso para porcentajes de datos faltantes cercanos al 80 %. El RMSE para SEMA esta fuera de la escala elegida en el gráfico, mientras que JEMA tiene curvas de RMSE planas para todos los porcentajes de datos faltantes.

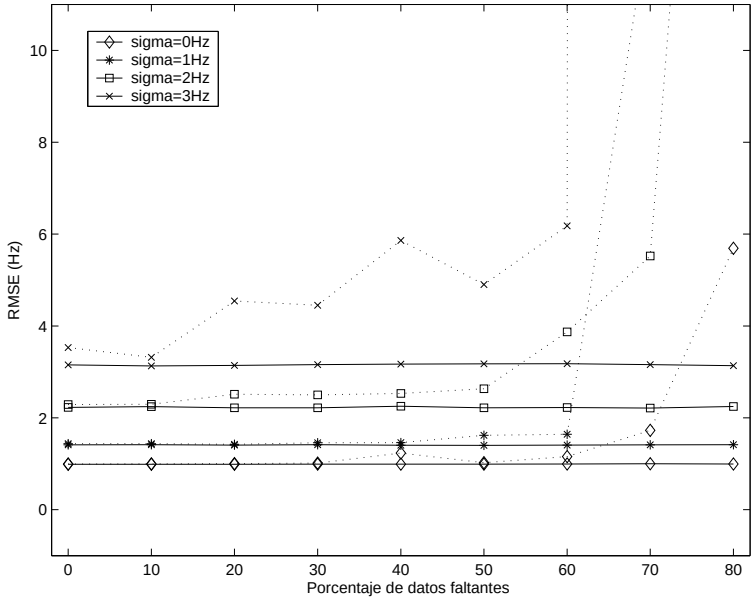


Figura 4.3: RMSE obtenido usando parametrización de Fujisaki para diferentes condiciones de ruido e información faltante en los datos de entrenamiento.

Estas simulaciones muestran la importancia de la correcta elección del enfoque de entrenamiento, tal como ocurre en este caso con JEMA.

4.2. Validación de JEMA para el modelado de la entonación

En este apartado se presentan los experimentos diseñados para analizar las mejoras proporcionadas por el enfoque de extracción y modelado conjunto (JEMA) con respecto a las propuestas que se pueden encontrar en la literatura, donde la extracción de parámetros y el modelado se realizan de forma independiente (SEMA).

Se han considerado cinco modelos de entonación:

- **Bézier SEMA (BAS)** [Agü04a]. En este caso la unidad prosódica es el grupo acentual. En un primer paso se extraen los parámetros de los contornos originales (coeficientes de Bézier). Luego, se construye un árbol que predice conjuntamente todas las componentes de los parámetros usando una regresión lineal sobre el vector completo de coeficientes de Bézier.
- **Bézier JEMA (BAJ).** Este modelo de entonación es idéntico al modelo anterior, pero el entrenamiento se realiza usando JEMA. El objetivo es comparar el rendimiento de modelos con igual cantidad de parámetros para observar las ventajas de JEMA sobre SEMA.

- **S-Bezier JEMA (BSJ) [Agü04a].** Modelo de entonación superposicional de Bézier, cuyas componentes son el grupo acentual y el grupo entonativo. Este modelo tiene una mayor flexibilidad que los otros modelos no superposicionales, y separa los efectos de la tendencia a la declinación asociada al grupo entonativo de los eventos asociados a la componente de grupo acentual.

- **Fujisaki SEMA (FS) [Agü04c].** El modelo de entonación de Fujisaki es superposicional. Existen dos componentes: acento (relacionada con el grupo acentual) y frase (relacionada con el grupo entonativo). En este caso, el modelo de entonación se genera extrayendo primero los parámetros y luego se entrenan dos árboles diferentes para la predicción de los mismos: árbol de comandos de acento y de comandos de frase. Este enfoque difiere de algunas propuestas en la literatura que usan un árbol para cada parámetro.

- **Fujisaki JEMA (FJ) [Agü04c].** Este modelo de entonación es idéntico al anterior, pero su entrenamiento se realiza usando JEMA.

4.2.1. Datos experimentales

Para los experimentos se decidió utilizar el corpus C1.1 (Apéndice B.4), el cual consiste de 1556 frases entonativas y 5643 grupos acentuales. Los límites de los grupos acentuales se determinan automáticamente mediante la información del acento léxico disponible en el corpus, utilizando la definición de la Sección 2.2.1: en el español el grupo acentual se encuentra constituido por una palabra acentuada y todas aquellas palabras no acentuadas que le preceden. Esta unidad ha sido utilizada por numerosos autores para describir los patrones de entonación del español a nivel local [Gar96, Alc98, Sos99, Esc02a].

En los experimentos se utilizaron las fronteras de frases entonativas que se encuentran etiquetadas en forma manual en el corpus. A pesar de que existen dos niveles de frases entonativas (grupos y cláusulas entonativas, que corresponderían a los niveles 3 y 4 de la capa *break index* de ToBI), en los experimentos se ha fusionado las dos clases en una.

Para cada uno de los modelos de entonación que están siendo evaluados se entrenó utilizando el enfoque de *20-fold cross validation* con el objeto de obtener mejores estadísticas del rendimiento de los sistemas. En consecuencia, los datos disponibles del corpus fueron divididos en 20 partes con cantidad similar de párrafos cada una.

El contorno de frecuencia fundamental fue estimado usando Praat [Boe] estableciendo mediante algunas observaciones previas el rango tonal de cada locutor, con el objeto de restringir los valores de frecuencia fundamental detectados. De esta manera, se reduce la posibilidad de la existencia de *pitch halving* o *pitch doubling*.

Las características utilizadas para el modelado de las frases entonativas incluyen el número de sílabas, palabras y grupos acentuales que las constituyen, e información sobre signos de puntuación en sus fronteras.

En el caso del grupo acentual se utilizó características relativas a su posición dentro del grupo entonativo, la posición de la sílaba acentuada, el número de sílabas y palabras que lo constituyen, e información sobre signos de puntuación en sus fronteras.

4.2.2. Resultados experimentales

En las Figuras 4.4 y 4.5 se pueden observar los resultados experimentales usando la voz femenina del corpus TC-STAR. Las Figuras 4.6 y 4.7 presentan los resultados experimentales usando la voz masculina del proyecto TC-STAR.

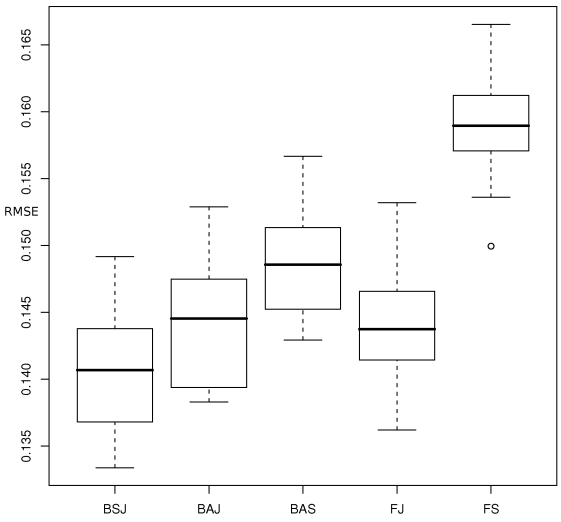


Figura 4.4: RMSE obtenido para los diversos modelos de entonación usando los datos de evaluación para el hablante femenino

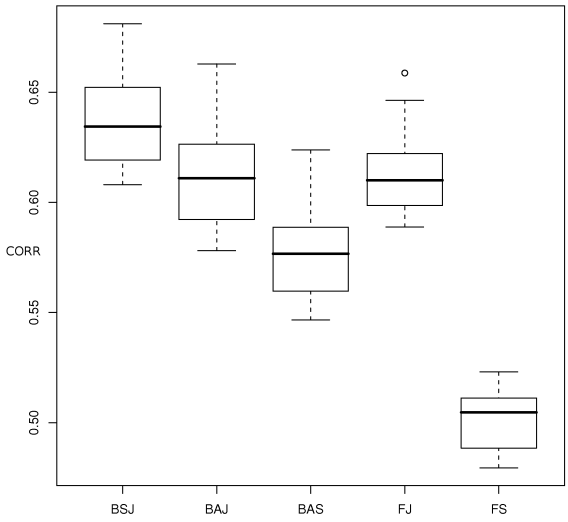


Figura 4.5: Correlación obtenida para los diversos modelos de entonación usando los datos de evaluación para el hablante femenino

El eje vertical corresponde al RMSE en escala logarítmica, siguiendo las recomendaciones de la literatura que indican que es más conveniente analizar el rendimiento de los modelos de entonación en la escala logarítmica que utilizando la escala lineal.

En cada una de las figuras se puede observar que todos los modelos que utilizan el enfoque JEMA (BSJ, BAJ y FJ) para el entrenamiento tienen un RMSE menor y una mayor correlación que aquellos modelos entrenados usando un enfoque de dos pasos (BAS y FS). Esta afirmación es estadísticamente significativa con una probabilidad $p < 0{,}01\,\%$.

El modelo de entonación superposicional con funciones de Bézier (BSJ) presenta los mejores resultados objetivos tanto en RMSE como en correlación. Mientras tanto, los otros modelos entrenados usando JEMA (BAJ y FJ) obtienen resultados similares entre sí, e inferiores a BSJ con probabilidad $p < 5\,\%$ y $p < 0{,}01\,\%$ respectivamente.

El modelo de entonación de Fujisaki tiene un mayor RMSE que el modelo superposicional basado en coeficientes de Bézier debido a las limitaciones de la representación matemática exponencial utilizada. Es esperable que la flexibilidad de la formulación del modelo de entonación que usa parámetros de Bézier (se han utilizado contornos de orden 4 tanto para la componente de frase como para la acentual) se refleje en resultados

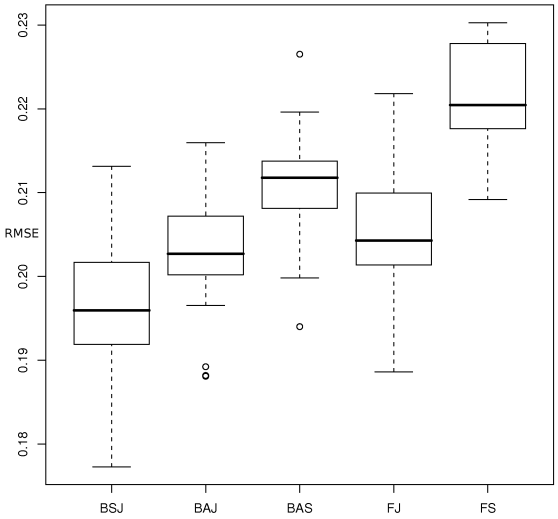


Figura 4.6: RMSE obtenido para los diversos modelos de entonación usando los datos de entrenamiento para el hablante masculino

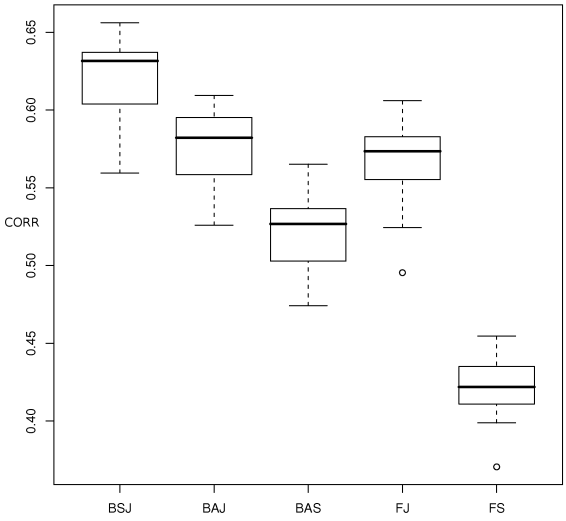


Figura 4.7: Correlación obtenida para los diversos modelos de entonación usando los datos de evaluación para el hablante masculino

objetivos mejores que los modelos que usan la parametrización propuesta por Fujisaki. Las funciones exponenciales tienen menores posibilidades de aproximar adecuadamente un contorno arbitrario de entonación que una representación polinómica.

La utilización de SEMA usando funciones de Bézier (BAS) produce un modelo que tiene un rendimiento ligeramente inferior a los otros, con excepción de FS.

Por último, la consistencia obtenida para el modelo de Fujisaki usando JEMA (FJ) se ve contrastada con los bajos resultados obtenidos por SEMA (FS) ($p < 0{,}01\,\%$). Este último modelo de entonación obtiene resultados muy inferiores a los otros modelos tanto en RMSE como en correlación.

La mejora en el modelo de entonación de Bezier es probable que no pueda ser percibida claramente por el oyente debido a que es pequeña. Sin embargo, en el caso del modelo de entonación de Fujisaki los contornos predichos usando JEMA posiblemente se percibirán claramente como mejores.

Dentro de las características más relevantes para el modelado del grupo acentual se encuentran su posición relativa dentro de la frase entonativa, los signos de puntuación, y el número de sílabas y palabras que lo constituyen.

En el modelado de los grupos entonativos, las características más importantes corresponden a los signos de puntuación, y al número de sílabas y grupos acentuales que los componen.

Los tiempos de entrenamiento para los 70 minutos de contornos de frecuencia funda-

mental son similares para todos los modelos de entonación excepto el modelo de Fujisaki entrenado con JEMA. El tiempo necesario para obtener los modelos de entonación fué: de 15 minutos para BSJ, 4 minutos para BAJ, 3 minutos para BAS y 6 minutos para FS. En el caso del modelo de entonación de Fujisaki entrenado con JEMA, el tiempo de entrenamiento medio es de 6 horas. Esta gran diferencia con respecto a los otros modelos se origina en la necesidad de utilizar algoritmos de gradiente para obtener la solución óptima de los parámetros, debido al carácter no lineal de la solución de las ecuaciones de optimización (Sección 3.1.4).

Evaluación subjetiva

Con el objeto de complementar los resultados objetivos, se llevó a cabo una evaluación subjetiva acerca de la naturalidad y la calidad de los distintos modelos de entonación.

Se utilizó PRAAT [Boe] para resintetizar los párrafos de los datos de evaluación usando el contorno predicho por cada modelo. La resíntesis preserva en gran medida la calidad del audio, siendo solamente afectada por la modificación del parámetro bajo evaluación.

Para ello, se le solicitó a 25 expertos evaluar la naturalidad con una puntuación entre 1 (completamente no natural) y 5 (completamente natural). También se les solicitó evaluar la calidad del audio entre 1 (baja calidad) y 5 (alta calidad).

Cada evaluador puntuó 12 párrafos correspondientes a 6 entonaciones: 5 modelos de entonación y voz real (no resintetizada). La voz real se incluye para validar la comprensión de la tarea por parte del evaluador. Para cada modelo se eligieron aleatoriamente dos párrafos entre los existentes en los datos de evaluación. Lo mismo se realizó para la voz real.

En la Figura 4.8 se puede observar los resultados de naturalidad de la evaluación subjetiva y en la Figura 4.9 los correspondientes a la calidad para cada uno de los modelos de entonación y la entonación real (R).

La entonación real (R) posee un MOS (*Mean Opinion Score*) claramente diferenciado de todos los modelos de entonación sujetos a evaluación. El modelado de la entonación no ha logrado imprimir en los evaluadores la percepción de completa naturalidad.

La naturalidad de los modelos de entonación superposicionales de Bézier y Fujisaki entrenados con el enfoque JEMA (BSJ y FJ, respectivamente) es ligeramente superior a la obtenida con BAJ y BAS.

El modelo de entonación de Fujisaki entrenado con el enfoque SEMA posee la naturalidad más baja, claramente diferenciada de todos los otros modelos.

La diferenciación entre el contorno real y los modelos de entonación también se observa en la calidad, principalmente debido a la manipulación acústica del algoritmo PSOLA. Los modelos de entonación BSJ, BAJ, BAS y FJ tienen igual mediana y ubicación del primer cuartil.

Por otra parte, el modelo de Fujisaki entrenado con SEMA tiene la calidad percibida más baja. Esto se debe a que el modelo de entonación FS posee el RMSE más alto, y por ello la manipulación acústica es más grande. Esto contribuye a una calificación más baja

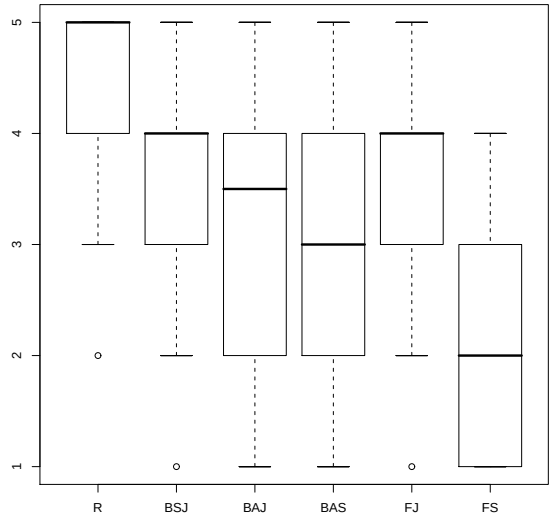


Figura 4.8: MOS de *naturalidad* obtenido para los diversos modelos de entonación

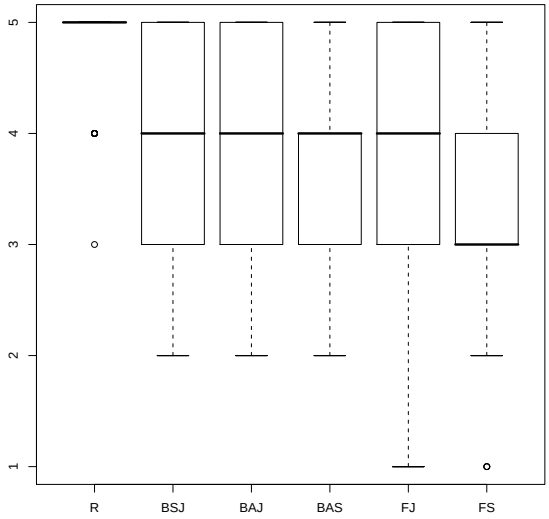


Figura 4.9: MOS de *calidad* obtenido para los diversos modelos de entonación

del audio resintetizado usando las curvas de entonación predichas por dicho modelo.

Un análisis usando el *Mann-Whitney-Wilcoxon test* con $p = 0{,}05$ (Tabla 4.2) revela que hay diferencias significativas entre las distribuciones de MOS del contorno real y todos los modelos, y entre todos los modelos y el modelo FS.

Modelo	**Naturalidad**	**Calidad**
R	4.63	4.78
BSJ	3.55	3.78
BAJ	3.26	3.60
BAS	3.05	3.44
FJ	3.5	3.81
FS	2.34	3.07

Tabla 4.1: MOS de naturalidad y calidad obtenido para los diversos modelos de entonación usando los datos de evaluación

	R	BSJ	BAJ	BAS	FJ	FS
R		■	■	■	■	■
BSJ			□	■	□	■
BAJ				□	□	■
BAS					■	■
FJ						■

Tabla 4.2: *Mann-Whitney-Wilcoxon test* de la evaluación subjetiva de la naturalidad de la entonación para los diversos modelos

Además, las diferencias entre BSJ y BAS, y FJ y BAS son significativas, demostrando una preferencia por la utilización de enfoques superposicionales entrenados con JEMA.

Tampoco existen preferencias significativas por la entonación de oraciones cortas en comparación a la de las oraciones más largas en los resultados de la evaluación subjetiva. En consecuencia, no puede inferirse que las oraciones de mayor duración tengan una entonación poco natural originada por el carácter repetitivo de los contornos tonales o la monotonía de la entonación.

4.3. Validación de JEMA para el modelado de la duración

Los experimentos sobre modelado de la duración segmental usando dos niveles (segmental y suprasegmental) se realizaron usando árboles de regresión. El objetivo de los experimentos era estudiar las ventajas del uso de dos niveles sobre un modelado tradicional de la duración segmental independiente de la duración silábica.

Se han considerado cinco modelos de duración:

- **Predicción de la duración segmental sin utilización de información contextual (DPR).** Este sistema base modela la duración segmental sin considerar el contexto fonético. Este modelo se incluyó como uno de los sistemas más simples de predicción de la duración segmental, y solamente incluye como información contextual la presencia de una pausa a continuación de un fonema.
- **Predicción de la duración segmental con utilización de información contextual (DP).** La duración segmental se predice utilizando tanto información fonética del segmento como del contexto. Este modelo mejora la información provista al modelo DPR para la predicción de la duración segmental, mediante la inclusión de características de los fonemas adyacentes. El mismo se incluye porque es otro de los sistemas clásicos para la predicción de la duración segmental.
- **Predicción de la duración segmental en base a la duración de la sílaba (DS) (Sección 3.2.2).** La duración segmental se predice con una fracción de la duración predicha de la sílaba. Este enfoque no considera la interacción de los dos niveles en el proceso de aprendizaje automático.
- **Predicción de la duración segmental en base a la duración de la sílaba optimizado usando duraciones relativas (DSO) (Sección 3.2.3).** La duración segmental se predice como una fracción de la duración predicha de la sílaba. Este enfoque considera la interacción de los dos niveles en el proceso de aprendizaje automático.
- **Predicción de la duración segmental en base a la duración de la sílaba optimizado usando duraciones relativas y absolutas (DSM) (Sección 3.2.4).** La duración segmental se predice como una fracción de la duración predicha de la sílaba, o bien de manera absoluta sin considerar la duración suprasegmental. Este enfoque considera la interacción de los dos niveles en el aprendizaje.

4.3.1. Datos experimentales.

Los experimentos se realizaron usando las voces base del Proyecto Europeo TC-STAR (ver Apéndice C). El corpus de estudio está compuesto de párrafos correspondientes al dominio parlamentario (C1.1).

Se dispone para el estudio de la duración de 18.603 sílabas, mientras que el número de fonemas es 43.800. La silabificación se obtuvo en forma automática a través del uso de un conjunto de reglas del español.

La segmentación de los fonemas en el audio es automática usando la transcripción fonética disponible, que fué corregida manualmente, y el sistema de reconocimiento del habla de la UPC: RAMSES [Bon98]. Mediante el entrenamiento de Modelos Ocultos de Markov (HMM) de semifonemas con contexto, se determinaron las fronteras de los fonemas utilizando alineamiento forzado.

La duración suprasegmental se modela usando árboles de regresión y un conjunto de parámetros que son considerados relevantes para la estimación de la duración de la sílaba, tales como:

- Posición de la sílaba con respecto a la pausa más cercana.
- Presencia de un acento en la sílaba.
- Secuencia de fonemas constituyentes de la sílaba.
- Punto de articulación de los fonemas constituyentes de la sílaba.
- Modo de fonación de los fonemas constituyentes de la sílaba.
- Tipo de fonemas constituyentes de la sílaba (consonantes o vocales).
- Sonoridad de los fonemas constituyentes de la sílaba (sonoros o sordos).
- Posición de la sílaba relativa al grupo entonativo.
- Posición de la sílaba relativa a la palabra.
- Número de sílabas que constituyen la palabra.
- Número de fonemas que constituyen la sílaba.

La duración segmental se modela con otro conjunto de parámetros relevantes a ella:

- Características articulatorias del fonema: punto de articulación, modo de fonación, tipo de fonema y sonoridad.
- Características articulatorias del fonema precedente.
- Características articulatorias del fonema subsiguiente.
- Posición dentro de la sílaba: onset, núcleo o coda.
- Posición de la sílaba con respecto a la pausa más cercana.
- Presencia de un acento en la sílaba.
- Secuencia de fonemas constituyentes de la sílaba.
- Punto de articulación de los fonemas constituyentes de la sílaba.
- Modo de fonación de los fonemas constituyentes de la sílaba.

- Tipo de fonemas constituyentes de la sílaba (consonantes o vocales).
- Sonoridad de los fonemas constituyentes de la sílaba (sonoros o sordos).
- Número de fonemas que constituyen la sílaba.

En el caso del modelo DPR no se ha incluido la información fonética contextual, con el objetivo de observar su importancia tanto objetiva como subjetiva.

Los datos experimentales fueron divididos en 20 partes con cantidad similar de párrafos cada una, con el objeto de utilizar el enfoque de *20-fold cross validation* para obtener mejores estadísticas del rendimiento de los sistemas.

Debido a que en los resultados experimentales se deseaba incluir una evaluación subjetiva de naturalidad y calidad, esto motivó la utilización del párrafo para dividir los datos de entrenamiento

4.3.2. Resultados experimentales.

En las Figuras 4.10 y 4.11 se pueden observar los resultados experimentales usando la voz femenina y masculina del proyecto TC-STAR.

El eje vertical corresponde al RMSE en milisegundos, siguiendo las recomendaciones de la literatura que indican que es conveniente analizar el rendimiento de los modelos de duración usando el RMSE utilizando la escala lineal.

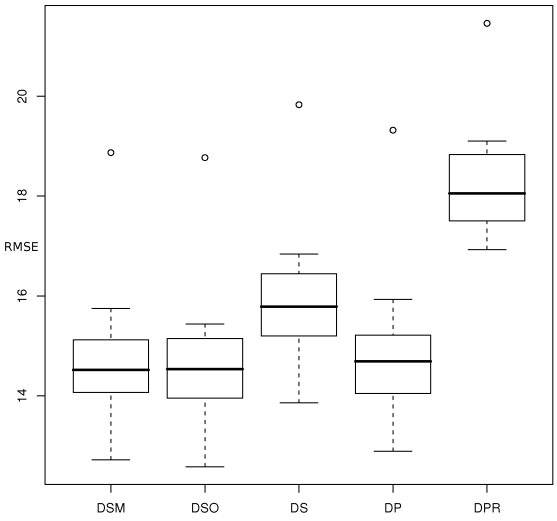


Figura 4.10: RMSE obtenido para los diversos modelos de duración: hablante femenino

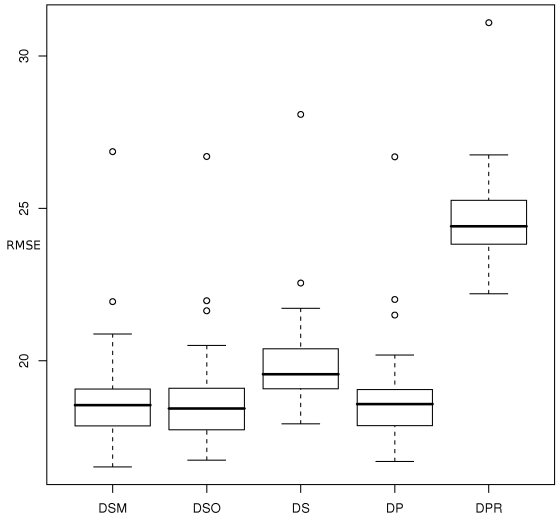


Figura 4.11: RMSE obtenido para los diversos modelos de duración: hablante masculino

En cada una de las figuras se puede observar que los modelos que utilizan un enfoque de dos niveles con entrenamiento conjunto (DSM y DSO) poseen rendimientos similares

al modelado segmental con información contextual (DP), y no hay diferencias estadísticamente significativas.

Además, también se observa la mayor consistencia en la predicción usando modelado conjunto (DSM y DSO) sobre el modelado usando dos niveles entrenados independientemente (DS). Las diferencias de los modelos DSM y DSO con respecto a DS son estadísticamente significativas con probabilidad $p < 0{,}03\,\%$ y $p < 0{,}02\,\%$, respectivamente. El entrenamiento por separado de cada uno de los niveles en DS conlleva a un incremento en el error de modelado.

El modelo de duración sin información contextual (DPR) posee un rendimiento claramente inferior a todos los otros modelos ($p < 0{,}01\,\%$), tal como era de esperar, debido a la ausencia de la importante información fonética del contexto.

Dentro de las características más relevantes en el modelado de la duración silábica se encuentran la cercanía a junturas terminales, la existencia de acento léxico, y el número de fonemas que componen la sílaba.

Con respecto a la información articulatoria de los fonemas que constituyen la sílaba, la sonoridad y el carácter consonántico o vocálico son las características más relevantes. Luego continúan en importancia el punto y el modo de articulación. En la posición menos destacada se encuentran el número de sílabas que constituyen la palabra y la identidad de los fonemas que la constituyen.

Las características más relevantes en el modelo DSM para la predicción de la duración segmental son la sonoridad, el carácter consonántico o vocálico, y el punto y modo de articulación tanto del fonema como de los fonemas adyacentes. La posición dentro de la sílaba y la cercanía de una juntura terminal también son relevantes.

Los tiempos de entrenamiento son similares para tres de los seis modelos de entonación. El tiempo necesario para obtener un modelo con la misma cantidad de datos de entrenamiento es de 3 minutos para DPR, 6 minutos para DP, y 10 minutos para DS.

En el caso de los algoritmos propuestos en esta tesis, los tiempos de entrenamiento son mayores, con un valor medio de 24 horas. Esta gran diferencia está motivada por la formulación matemática más compleja para estimar la duración segmental y silábica, lo cual hace necesario resolver ecuaciones no lineales mediante un método iterativo.

Evaluación subjetiva

Con el objeto de complementar los resultados objetivos, se llevó a cabo una evaluación subjetiva acerca de la naturalidad de la fluidez y la calidad de los distintos modelos de duración.

Se utilizó PRAAT [Boe] para resintetizar los párrafos de los datos de evaluación usando el duración predicha por cada modelo. La resíntesis preserva en gran medida la calidad del audio, siendo solamente afectado por la modificación del parámetro bajo evaluación.

Para ello, se le solicitó a 25 personas evaluar la naturalidad con una puntuación entre 1 (fluidez completamente no natural) y 5 (fluidez completamente natural). También se les solicitó evaluar la calidad del audio: 1 (baja calidad) y 5 (alta calidad).

Cada evaluador puntuó 12 párrafos correspondientes a 6 casos: 5 modelos de duración y voz real (no resintetizada). La voz real se incluye para validar la comprensión de la tarea por parte del evaluador. Para cada modelo se eligieron aleatoriamente dos párrafos entre los existentes en los datos de evaluación. Lo mismo se realizó para la voz real.

En las Figuras 4.12 y 4.13 se puede observar los resultados de la evaluación objetiva para cada uno de los modelos de duración y la voz real (R).

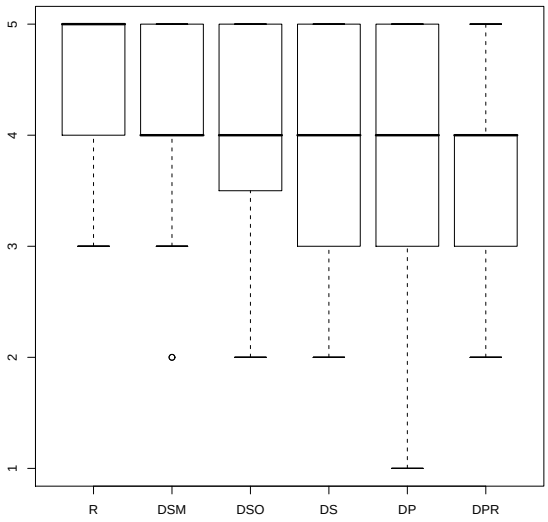


Figura 4.12: MOS de naturalidad obtenido para los diversos modelos de duración usando los datos de evaluación

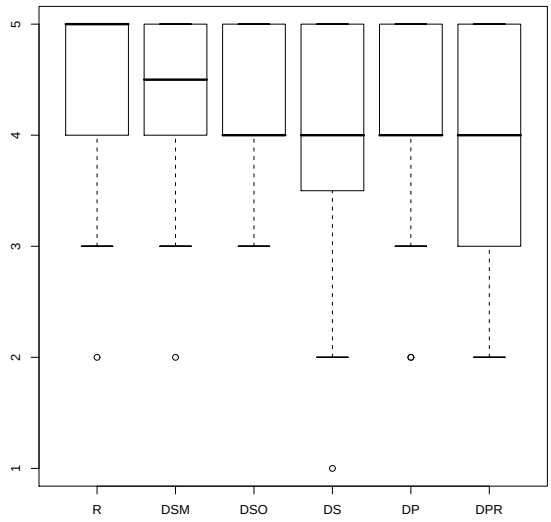


Figura 4.13: MOS de calidad obtenido para los diversos modelos de duración usando los datos de evaluación

La voz real (R) puede ser claramente diferenciada tanto en naturalidad como en calidad por parte de todos los evaluadores, recibiendo solamente en casos aislados una puntuación inferior a 5.

Uno de los modelos de duración propuestos (DSM) posee una distribución de la puntuación ligeramente superior al resto de los modelos. El mismo posee el 50 % de las puntuaciones entre 4 y 5, mientras que los otros modelos poseen el 50 % de las puntuaciones entre 3 y 5.

Por lo observado en las distribuciones de los cuartiles, no existen diferencias entre DS y DP, siendo DSO ligeramente superior a ellos.

El modelo de duración sin información contextual ha sido claramente diferenciado de los otros, recibiendo el 50 % de las puntuaciones entre 3 y 4.

Un análisis usando el *Mann-Whitney-Wilcoxon test* (Tabla 4.4) revela que solamente hay diferencias significativas entre la distribución de puntuaciones de la voz real y los otros modelos, y entre el modelo DSM y DPR.

Debido a que solamente existen diferencias significativas entre el modelo DPR y el

modelo DSM, se puede concluir que se evidencian problemas por parte de los evaluadores para determinar diferencias entre los sistemas.

Modelo	Naturalidad	Calidad
R	4.59	4.65
DSM	4.06	4.28
DSO	4.06	4.25
DS	3.87	4.00
DP	4.00	4.09
DPR	3.62	4.03

Tabla 4.3: MOS de naturalidad y calidad obtenido para los diversos modelos de duración usando los datos de evaluación

	R	DSM	DSO	DS	DP	DPR
R		■	■	■	■	■
DSM			□	□	□	■
DSO				□	□	□
DS					□	□
DP						□

Tabla 4.4: *Mann-Whitney-Wilcoxon test* de la evaluación subjetiva de la naturalidad de la duración para los diversos modelos

4.4. Experimentos sobre modelado de junturas terminales

En los experimentos para la evaluación de los modelos de junturas terminales se usaron los tres métodos mencionados en la Sección 2.4, tanto con palabras como con grupos acentuales, con el objeto de estudiar sus fortalezas y debilidades utilizando las mismas condiciones de experimentación:

- **CART**: Arbol de clasificación usando tres características: ventana de POS, puntuación y distancia de la última juntura terminal.
- **CART+LM**: Arbol de clasificación usando dos características (ventana de POS y puntuación) y un modelo de lenguaje de junturas terminales.
- **FST**: Transductor de estados finitos que usa etiquetas de POS y puntuación.
- **CART (AG)**. Idem a CART pero usando grupos acentuales en lugar de palabras.
- **CART+LM (AG)**. Idem a CART+LM pero usando grupos acentuales en lugar de palabras.
- **FST (AG)**. Idem a FST pero usando grupos acentuales en lugar de palabras.

4.4.1. Datos experimentales.

Los experimentos se realizaron usando los párrafos correspondientes al dominio parlamentario (C1.1), donde se encuentran 1556 junturas terminales en un total de 9337 palabras y 221 párrafos.

Las junturas terminales fueron etiquetadas manualmente utilizando dos niveles, y de ellas el 40 % no coincide con signos de puntuación. En esto experimentos no se hizo distinción entre ambos tipos de junturas terminales: grupo entonativo y cláusula entonativa.

La predicción de las junturas terminales usando árboles de clasificación (CART y CART+LM, tanto para palabras como grupos acentuales) utilizó un conjunto de características que son relevantes:

- Ventana de cinco etiquetas morfosintácticas, tres anteriores a la posición evaluada y dos posteriores a ella.
- Tipo de signo de puntuación en la posición.
- Distancia en sílabas y palabras a la última juntura terminal.
- Distancia en sílabas y palabras desde el signo de puntuación.

Los datos experimentales fueron divididos en 20 partes con cantidad similar de párrafos cada una, con el objeto de utilizar el enfoque de *20-fold cross validation* para obtener mejores estadísticas del rendimiento de los sistemas.

4.4.2. Resultados experimentales.

Para comparar los diferentes algoritmos de predicción de junturas terminales se usaron cuatro medidas, las cuales se encuentra en la literatura acerca de la evaluación de los modelos de junturas terminales (ver Tabla 4.5):

- **Exactitud**: Porcentaje de junturas terminales (J) y no-junturas terminales ($\overline{J}$) colocadas correctamente: $\frac{a+d}{a+b+c+d}$.
- **Precision**: Porcentaje de junturas terminales correctas sobre el total de junturas terminales predichas: $\frac{a}{a+b}$.
- **Cobertura**: Porcentaje de las junturas terminales predichas correctamente.: $\frac{a}{a+c}$.
- **F-measure**: Media armónica de precisión y recall: $\frac{2.precision.recall}{precision+recall}$.

	J	$\overline{J}$
J	(a)verdadero-positivo	(b)falso-positivo
$\overline{J}$	(c)falso-negativo	(d)verdadero-negativo

Tabla 4.5: Tabla de confusión

En las Tablas 4.6 y 4.7 se puede ver los resultados utilizando las distintas medidas para cada uno de los seis métodos. La exactitud es la misma para todos ellos, alrededor del 93 % para el hablante femenino y 88 % para el hablante masculino. Estos valores son buenos debido a que la exactitud del sistema básico (que es aquel que predice siempre que no hay junturas terminales) es del 83 % para el hablante femenino y 78 % para el hablante masculino. Sin embargo, es necesario prestar atención a otros parámetros tales como recall y precisión para tener una idea acerca de la calidad de las junturas terminales.

Medida	**CART**	**CART+LM**	**FST**	**CART**$_{(AG)}$	**CART+LM**$_{(AG)}$	**FST**$_{(AG)}$
Exactitud	93.14	94.08	93.02	93.02	93.92	93.07
Precision	89.32	86.98	86.66	88.85	87.91	87.01
Recall	67.99	76.86	72.25	67.54	74.79	71.35
F-Measure	77.09	81.50	78.59	76.59	80.67	78.28

Tabla 4.6: Resultados experimentales del modelado de las junturas terminales del hablante femenino.

Medida	**CART**	**CART+LM**	**FST**	**CART**$_{(AG)}$	**CART+LM**$_{(AG)}$	**FST**$_{(AG)}$
Exactitud	88.23	89.24	88.52	87.93	89.32	88.70
Precision	78.34	76.49	78.91	76.96	77.82	77.14
Recall	65.16	74.80	68.95	64.68	73.03	71.99
F-Measure	71.04	75.53	73.55	70.23	75.26	74.43

Tabla 4.7: Resultados experimentales del modelado de las junturas terminales del hablante masculino.

Ninguno de los sistemas tiene una precisión mejor que los otros sistemas, y su valor es cercano al 87 % para el hablante femenino y 78 % para el hablante masculino. Esto significa un alto porcentaje de junturas terminales predichas correctamente sobre el total de junturas predichas para el hablante femenino. El resultado es importante, debido a que es necesario ubicar junturas terminales en los lugares correctos con poco errores.

El recall solamente mide cuantas junturas terminales predichas son correctas sobre el total. Las Tablas 4.6 y 4.7 muestran que el sistema con mejor recall es CART+LM, tanto para palabras como para grupos acentuales.

F-Measure es la media armónica de la precisión y el recall. De esta manera podemos combinar dos de los factores que son importantes en uno. Los resultados que se muestran en las tablas indican que el mejor compromiso entre los dos parámetros lo ofrece CART+LM para el hablante femenino y el hablante masculino.

La característica más relevante ha sido la distancia en palabras hasta el próximo signo de puntuación, seguida por las etiquetas morfosintácticas de la palabra previa a la posición evaluada y las dos etiquetas siguientes. Luego continúa en importancia la distancia en palabras desde la última juntura terminal. Al final de esta lista se encuentran la etiqueta morfosintáctica de la palabra previa a la posición evaluada y la distancia en sílabas desde la última juntura terminal.

Una revisión manual de las junturas terminales predichas por los diferentes algoritmos muestra que dado que hay diferentes formas de pronunciar las frases se han contado como errores algunas decisiones que de hecho eran correctas. La exactitud, precisión, recall y F-measure son solamente cotas inferiores al rendimiento real de los sistemas.

La importancia de la falta de comprensión del lenguaje natural por parte de los algoritmos también se observa en la colocación de junturas terminales en lugares incorrectos, debido a la existencia de expresiones u oraciones subordinadas complejas. A continuación se muestran dos ejemplos:

"... incluso agresiva a veces (juntura terminal) como respuesta a nuestras dudas (juntura terminal no predicha) relativas a la situación ..."

"... en los que se basa esa misma competencia (juntura terminal) y es como si serraran la rama (juntura terminal no predicha) sobre la que se apoyan ..."

4.5. Conclusiones

En este capítulo se han presentado los experimentos realizados con algoritmos de modelado de distintos parámetros prosódicos: entonación, duración segmental y junturas terminales.

Se han comparado modelos existentes en la literatura con los propuestos en esta tesis a lo largo del Capítulo 3, con el objeto de analizar sus fortalezas y el grado de progreso alcanzado. En todos los casos se ha incluido tanto una medición objetiva como subjetiva de las diferencias entre los diversos modelos bajo estudio.

4.5.1. Entonación

A lo largo de las dos primeras secciones de este capítulo se hizo una presentación de los resultados experimentales de los diferentes modelos de entonación (S-Bezier, Bezier y Fujisaki) utilizando dos enfoques de entrenamiento: SEMA y JEMA.

En primer lugar se utilizaron contornos generados artificialmente (Sección 4.1) con el objeto de tener condiciones experimentales controladas en lo relativo a la complejidad de los movimientos tonales y a la disponibilidad del total de las características lingüísticas. Como resultado de esta configuración experimental es posible encontrar los contornos pertenecientes a cada clase de movimiento tonal debido a que se ajustaban a la forma de la parametrización, y las características lingüísticas permitían la completa separabilidad entre clases.

Los resultados experimentales demostraron que el enfoque JEMA es superior a SEMA para diferentes niveles de ruido (simulando microprosodia y errores de estimación) e información faltante en los contornos de entonación (debido a segmentos sordos). Las fortalezas de JEMA se reflejaron tanto en el modelo de entonación que utiliza curvas de Bezier como en el modelo de entonación de Fujisaki.

En la Sección 4.2 se hicieron experimentos usando contornos reales con los diversos modelos de entonación (S-Bezier, Bezier y Fujisaki) tanto con el enfoque SEMA como JEMA. Los datos utilizados corresponden a dos locutores, de sexo femenino y masculino, grabados en el marco del proyecto europeo TC-STAR.

También en este caso los resultados experimentales permitieron corroborar la superioridad del enfoque de entrenamiento conjunto JEMA sobre el enfoque SEMA. Los resultados objetivos señalan una mejora en el error cuadrático medio y en la correlación de la contornos estimados con respecto a los contornos reales.

Con el fin de obtener también una medida subjetiva de la aceptación, se llevó a cabo

una evaluación perceptual de naturalidad y calidad de los contornos generados por cada uno de los modelos de entonación. En la evaluación se incluyeron contornos reales para tener una medida de la capacidad de los evaluadores para identificar naturalidad y calidad.

Los resultados subjetivos revelan una preferencia por los modelos entrenados usando el enfoque JEMA. Estas diferencias demostraron ser estadísticamente significativas ($p = 0{,}05$) en muchos de los casos en que un modelo entrenado con JEMA fue comparado con un modelo entrenado usando SEMA, tal como indica el *Mann-Whitney-Wilcoxon test* realizado.

4.5.2. Duración

En la Sección 4.3 se presentaron los resultados de la evaluación experimental sobre el modelado de la duración segmental. Se han estudiado cinco modelos de duración, donde dos de ellos son puramente segmentales (DP y DPR), y los otros tres combinan el modelado segmental y suprasegmental (DS, DSO y DSM).

Todos los algoritmos demostraron pocas diferencias a nivel objetivo utilizando el error cuadrático medio como medida del error cometido entre la duración segmental predicha y la real. El modelo DPR es claramente inferior a todos los otros, y DS tiene un error cuadrático medio ligeramente mayor que los otros modelos: DP, DSO y DSM.

La evaluación objetiva de naturalidad y calidad revela también pequeñas diferencias entre los algoritmos DS, DP, DSO y DSM, que no resultan estadísticamente significativas para $p = 0{,}05$ usando el *Mann-Whitney-Wilcoxon test*.

Los algoritmos propuestos presentan una mejora consistente en cada uno de los conjuntos de evaluación utilizando *20-fold cross validation*. Sin embargo, estas diferencias son demasiado pequeñas para ser estadísticamente significativas en el conjunto de veinte experimentos realizados.

Esta cantidad limitada de datos de análisis también se ve reflejado en la evaluación subjetiva, donde solamente uno de los algoritmos propuestos (DSM) ha resultado con diferencias estadísticamente significativas con respecto a los otros modelos.

Es importante aclarar que una mayor cantidad de datos no necesariamente implicará una mejor estimación de la duración, ya que la ausencia de cierta información no obtenible en forma automática del texto hará que algunas clases no sean separables. Además, es necesario indicar que cierta falta de precisión en la segmentación automática puede dificultar la tarea de la estimación de la duración segmental, y esto verse reflejado en una degradación de los resultados obtenidos.

4.5.3. Junturas terminales

La Sección 4.4 incluyó los experimentos realizados para el modelado de junturas terminales con diversos algoritmos: CART, CART+LM y FST. Para cada uno de ellos se experimentó tanto con el uso de palabras como con grupos acentuales.

Los experimentos realizados revelan la ventaja de la utilización de modelos de lenguaje

a través de n-gramas sobre el algoritmo más simple que predice junturas usando CART. Tanto en el modelado usando palabras como grupos acentuales, CART+LM y FST resultaron superiores a la utilización de árboles de clasificación en forma aislada.

Además, en todos los casos CART+LM resulta superior a FST debido a la posibilidad de utilizar información contextual más compleja a través de la probabilidad modelada con el árbol de clasificación, tales como etiquetas morfosintácticas adyacentes y la distancia a signos de puntuación.

En los resultados no se observan grandes beneficios en la utilización del grupo acentual para la predicción de junturas terminales en lugar de la palabra. La única ventaja resulta en una disminución de la cantidad de decisiones que se deben tomar, reduciendo la carga computacional. Es de destacar que la ganancia es mínima debido a que la carga computacional es ínfima.

El análisis manual de las junturas predichas permitió verificar que las medidas objetivas utilizadas para medir el rendimiento de cada uno de los algoritmos son solamente una cota inferior al rendimiento real. El uso de una sola referencia tiende a ofrecer una medida pesimista del rendimiento de los modelos. Esto último es un problema común también observado en otros campos, como es el caso de la traducción automática, donde en ocasiones se utilizan múltiples referencias para medir la calidad de los sistemas en forma más precisa [Pap02].

En los Capítulos 3 y 4 se han propuesto y evaluado diversos modelos para la generación automática de la prosodia en los conversores texto a voz: entonación, duración segmental y junturas terminales. En el Capítulo 5 propondremos una extensión de estos modelos para su aplicación en el contexto de la traducción voz a voz, con la intención de aprovechar la información del idioma y el hablante de la lengua origen para mejorar la naturalidad y la expresividad de la prosodia generada en el conversor texto a voz de la lengua destino.

Capítulo 5

Transferencia de la prosodia en la traducción oral

En este capítulo se explicarán algoritmos para el aprovechamiento de las múltiples fuentes de información en un sistema de traducción voz a voz con el fin de mejorar la calidad de la conversión texto a voz. Entre estas fuentes se pueden mencionar: transcripción ortográfica del locutor fuente, fronteras de palabras y fonemas, pausas, información sobre alineamiento de las palabras del idioma fuente y destino, curva de frecuencia fundamental del locutor fuente e información sobre puntuación obtenida del ASR. El objetivo general es contribuir a convertir la traducción voz a voz en completa, abarcando desde el contenido, su forma de expresarlo a través de la prosodia, hasta llegar a incluir la identidad de la voz del hablante origen en la salida del sintetizador de voz, usando técnicas de conversión de voz.

Primeramente, en la Sección 5.1 se tratarán las limitaciones existentes para la generación de una prosodia natural y expresiva en un sistema de conversión texto a voz. Luego, en la Sección 5.2 se tratarán alternativas para la mejora de la generación de la prosodia en el marco de la traducción voz a voz, aprovechando las nuevas fuentes de información existentes en estos sistemas.

En las siguientes secciones del capítulo se explicarán los diferentes algoritmos propuestos. En la Sección 5.3 se detallará un algoritmo de transferencia de la entonación. La Sección 5.4 trata aspectos de sincronización y su relación con la duración segmental y las pausas. Finalmente, la Sección 5.5 contiene algunas propuestas para el uso de las pausas del idioma fuente para la mejora de la predicción de pausas en la conversión texto a voz.

5.1. Limitaciones para la generación de la prosodia en un sistema de conversión texto a voz

Como se mencionó al comienzo del Capítulo 1, la traducción voz a voz automática tiene como objetivo la traducción de la voz en un idioma y su reproducción en otro idioma en forma automática y sin la necesidad de intervención humana. Esto constituye un paso

adelante con respecto a la traducción texto a texto, debido a que se realiza utilizando el habla, mediante la inclusión en el proceso de áreas tales como el reconocimiento automático de voz (ASR) y la generación de voz por computadora (TTS).

Entre los objetivos de la tesis se encuentran el desarrollo de nuevos algoritmos para el entrenamiento de modelos de generación de prosodia para la conversión texto a voz, y su aplicación en el marco de la traducción voz a voz. Para ello se investiga la posibilidad de mejorar la naturalidad y expresividad de la conversión texto a voz utilizando la prosodia del hablante fuente disponible en el proceso de traducción voz a voz como información adicional.

A continuación se explicará la importancia de ciertos aspectos de la prosodia y la dificulad para ser generadas por un conversor texto a voz sin información adicional al texto, como la disponible en la traducción voz a voz.

En la introducción del capítulo 2 se explicó la importancia y el uso que se hace de la prosodia para estructurar el habla y el discurso a través de diferentes recursos acústicos, tales como la entonación, ritmo, intensidad, pausas, etc. En general, se puede afirmar que no es posible entender una oración sin el uso de los recursos prosódicos debido a la gran cantidad de información que proporcionan.

Quilis en su "Tratado de Fonología y fonética española" [Qui93] presenta diferentes funciones de la prosodia: distintiva, integradora, delimitativa, constrastiva y semántica. El correcto uso de las mismas en el habla sintetizada contribuye a una mejor opinión en términos de calidad y naturalidad por parte de los usuarios finales [Ant03]. Estas funciones de la prosodia son:

- La *función distintiva* permite entender el significado de la oración de acuerdo a las características prosódicas que son empleadas por el hablante. Por ejemplo, dependiendo de la pendiente del contorno de frecuencia fundamental al final de una oración, podemos diferenciar una oración declarativa de una interrogativa.
- La *función integradora* agrupa las unidades sin acento dentro de una unidad acentuada. De esta manera se puede entender el significado de la oración. Por ejemplo, la oración "estabarriendo" tiene un significado distinto de acuerdo a los acentos y a las junturas terminales: "está barriendo", "ésta va riendo" o "estaba riendo".
- La *función delimitativa* divide a la oración en unidades más pequeñas debido a razones fisiológicas (necesidad de respirar para continuar hablando) o razones gramaticales y lingüísticas (distribución de la información que hace al mensaje más entendible).
- La *función contrastiva* contribuye a mantener la atención del oyente en las partes importantes de la oración y a evitar la monotonía. Para ello se usan recursos tales como variaciones del ritmo o pausas largas para atraer la atención sobre una porción del discurso, separación en sílabas de la palabra a enfatizar, etc.
- La *función semántica* de la prosodia introduce en el mensaje hablado información adicional que clarifica el significado planeado por el locutor. Adicionalmente, cuando el significado de la prosodia y del mensaje son contradictorios, la prosodia se usa como

un primer indicador del significado real. La prosodia tiene poder para desambiguar entre diferentes significados de la oración.

Observando las múltiples funciones de la prosodia se puede inferir que su modelado no es una tarea fácil. Los sistemas de conversión texto a voz actuales tienen capacidades limitadas en los algoritmos de procesamiento de lenguaje natural utilizados para analizar una oración. Ello impide que la prosodia generada sea de una alta calidad y naturalidad.

Desde el punto de vista de la función distintiva encontramos que a través de la prosodia podemos diferenciar entre enunciados interrogativos y declarativos. La generación de la prosodia en conversión texto a voz debe analizar la existencia en el texto de signos de interrogación o pronombres interrogativos para decidir si el enunciado a sintetizar es declarativo o interrogativo. Si dicha información no se encuentra disponible resultará dificil tomar una decisión, ya que no sería posible conocer la intención del enunciado, tal como ocurre en el siguiente ejemplo:

Sin puntuación: *El presidente del gobierno ha dicho esta mañana que subirá los impuestos*
Con puntuación: *El presidente del gobierno, ¿ha dicho esta mañana que subirá los impuestos?*

En lo relativo a la función integradora los conversores texto a voz presentan limitaciones para enlazar porciones del discurso en unidades tales como grupos acentuales y entonativos. En algunas ocasiones es posible encontrar la frontera de un contorno melódico que preserve el sentido del mensaje, por ejemplo, a través de la presencia de signos de puntuación. Sin embargo, la decisión en porciones del texto sin signo de puntuación todavía resulta una tarea propensa a errores debido a que los ordenadores no pueden comprender el texto y decidir el sentido del mensaje. Estos mismos problemas ocurren con la función delimitadora, que tal como indican algunos autores, es complementaria de la función integradora. Por ejemplo, una pausa permitirá delimitar y diferenciar unidades de sentido o grupos fónicos que integrados perderían su sentido, como se demuestra en el siguiente ejemplo:

Mientras <pausa> el presidente del gobierno se ha referido al terrorismo.
Mientras el presidente del gobierno se ha referido al terrorismo.

En la primera oración la pausa permite delimitar las dos partes del enunciado, resultando una oración declarativa de la forma adecuada. En cambio, la ausencia de pausa en la segunda oración provoca que el oyente espere que la oración continúe luego de la palabra terrorismo, resultando en una confusión cuando el locutor hace una pausa luego de la palabra terrorismo debido al final de oración. Un conversor texto a voz tendrá dificultades para predecir dicha pausa, excepto que luego de la palabra “Mientras” exista una coma.

El **énfasis** se encuentra entre una de las características más difíciles de obtener y que es de extrema importancia para el modelado de la prosodia, ya que cumple una función contrastiva. Cada día enfatizamos partes del discurso cuando hablamos con el

objetivo de atraer la atención del oyente en ciertas partes del mismo, o bien para diferenciar información nueva de aquella ya mencionada. La ausencia de énfasis hará que nuestro modo de hablar sea juzgado como monótono: cada palabra tendrá la misma importancia que las otras. En consecuencia, el discurso será difícil de entender por el oyente.

En este aspecto, la ausencia de conocimiento acerca del mundo real por parte de los ordenadores es una limitación importante para el modelado de la prosodia. Por ejemplo, en la oración "Un turista fue detenido en el aeropuerto con explosivos en su zapato", el énfasis en la palabra "explosivos" no es el mismo que en la oración "Una tonelada de explosivos fue usada para demoler el viejo edificio de la estación de trenes". Los explosivos en un aeropuerto son más importantes que cuando son usados para hacer una demolición controlada de un viejo edificio. Es necesario el conocimiento acerca del mundo real para captar esa diferencia y generar una prosodia correcta.

En algunos casos, el hablante puede indicar su **intención** usando algunos recursos acústicos que introducen pequeños cambios en el significado del mensaje contenido en las palabras. Si la oración "lo compartiré contigo" se pronuncia con un ritmo lento y algunas pausas entre palabras, podemos convertir una afirmación en una duda. De esa manera expresaríamos que no estamos seguros que queremos compartirlo. Esta función semántica de la entonación que permite cambiar el significado de las palabras tiene una gran importancia para la generación de habla expresiva en los conversores texto a voz. Existen numerosos trabajos sobre el tema que proporcionan estudios sobre rasgos prosódicos de los distintos estados emocionales, como así también de la actitud del hablante con respecto al oyente y al contenido del mensaje [Sch09].

Además de las funciones enumeradas por Quilis, la prosodia también es importante para indicar las distintas **partes de un discurso**: narración, descripción, argumento, explicación, diálogo, etc, proporcionando una estructura al mismo y facilitando su comprensión. La ausencia de los marcadores prosódicos dificultan el entendimiento y disminuyen la capacidad de captar la atención del oyente durante la locución.

Finalmente, también es de importancia destacar que el contenido del mensaje determina el **estilo** con el que debe ser transmitido al oyente de acuerdo a algunas reglas acordadas por la sociedad. Por ejemplo, las noticias sobre una guerra no se leen de la misma manera que las noticias de deportes. Incluso las noticias de deportes se leen de manera diferente de acuerdo a su contenido. La noticia sobre la lesión de un jugador se lee con inquietud mientras que las noticias de resultados deportivos tendrán un estilo más vivaz. Además, en algunos programas de prensa partidaria las noticias sobre lesiones de jugadores rivales se pueden leer de manera alegre ya que beneficia a los resultados del otro equipo. Por ejemplo, una lesión de Ronaldo sería una buena noticia para algunos fanáticos del Barcelona.

En resumen, la lectura de textos necesita funciones cognitivas superiores que están lejos de poder ser realizadas por las computadoras debido a que carecen tanto de información sobre el mundo como de los diferentes rasgos de la psiquis humana (por ejemplo: emociones, intención, complejos, etc.). Las computadoras no pueden entender ni tener una opinión sobre lo que leen. Esto constituye un limite superior a la calidad alcanzable por cualquier sistema de conversión texto a voz.

Existen algunas propuestas en la literatura con el objeto de mejorar la calidad de

la síntesis introduciendo etiquetas para enriquecer la prosodia. En este marco existe un conjunto de iniciativas de estandarización usando XML, tales como SABLE [Spr98b], SSML [Bur04], JSML [Hun00], EML [EML], etc. El uso de etiquetas es importante en sistemas que generan voz de una manera muy controlada, como sucede con los sistemas concepto-a-voz [Piw02].

La aplicación de estas etiquetas se pueden ver en diversas aplicaciones, tales como:

- **Interfaces web de conversión texto a voz**, como es el caso de AT&T Labs Natural Voices (SSML) [ATTSit], ATalker (SABLE y SSML) [ATalke], Acapela (SSML y JSML) [Acapel], o el sintetizador de la UPC, UPCTTS (SABLE) [UPCTTS].
- **Plataformas robóticas**: plataforma robótica B21r con conversor texto a voz OpenMary (SSML) [Roe06], el robot PeopleBot (SSML) [Che08], o el robot para museos del proyecto INDIGO (SSML y JSML) [Vog08].
- **Telefonía**: RealSpeak de Telecom (SSML) [Tel05], o el sistema Say It Smart de Cisco (VoiceXML) [Cis07]
- **Plataformas de diálogos**: Gemini (SSML) [Cór04], o VoxNauta (SSML) [Gir09].

5.2. Generación de la prosodia en un sistema de traducción voz a voz

En un sistema de traducción voz a voz, el módulo de conversión texto a voz tiene más información que cuando se posee únicamente el texto, que es el caso cuando se utiliza un conversor texto a voz como un componente aislado. Tanto la voz del hablante fuente como las salidas de los sistemas de reconocimiento automático del habla y traducción automática son fuentes adicionales de información que pueden proporcionar importantes indicios para generar la prosodia de la conversión texto a voz, tal como se muestra en la Figura 5.1.

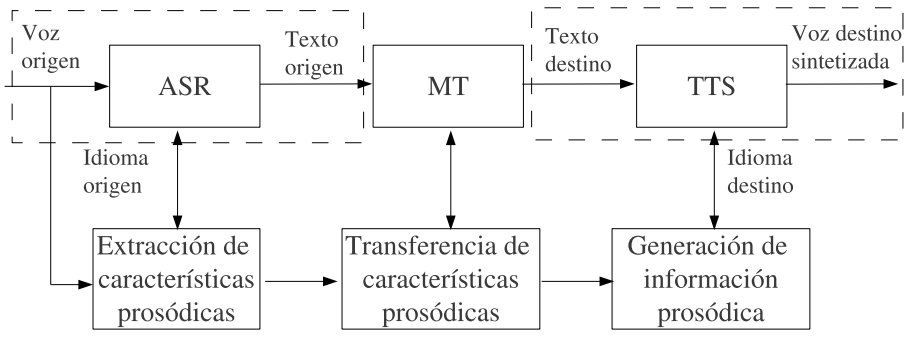


Figura 5.1: Esquema de generación de la prosodia utilizando la voz fuente.

En esta tesis se proponen una serie de algoritmos para contrarrestar las carencias de los conversores texto a voz en el procesamiento del lenguaje natural mencionadas en la sección anterior. Para ello se desarrollan una serie de modelos que usan tanto la información

acústica del hablante fuente como del texto traducido para la generación de la prosodia en el marco de la traducción voz a voz.

Los algoritmos desarrollados serán evaluados principalmente en la traducción voz a voz del idioma inglés al español, y viceversa. A pesar de que claramente los idiomas ingles y español tienen entonaciones distintas, la hipótesis que se desprende de la introducción del capítulo es que para reforzar una información ambos pueden llegar a utilizar la prosodia. Por ejemplo, una pregunta en un idioma tendrá un contorno inicial distinto que en el otro, pero ambos se auxilian en la función distintiva de la prosodia para indicar un enunciado interrogativo. Esto mismo ocurrirá con otros aspectos del discurso, tales como el énfasis (función contrastiva), las emociones, la intención, el estilo, etc.

En general podemos decir que las distintas funciones de la prosodia se encuentran representadas en ambos idiomas en diversas manifestaciones acústicas, tales como la entonación, el ritmo, las pausas, las junturas terminales, etc. El principal objetivo de este capítulo de la tesis es estudiar algoritmos que permitan aprovechar las manifestaciones acústicas en el idioma origen para enriquecer la prosodia en el idioma destino.

Una aproximación posible para la generación de la prosodia en el idioma destino dadas ciertas manifestaciones acústicas del idioma origen (tales como la entonación, el ritmo, las pausas, las junturas terminales, etc) consistiría en clasificar distintas porciones del texto de acuerdo a un conjunto de etiquetas predeterminadas que indicarían sentimientos, énfasis, contraste, etc. Un ejemplo de sistema de conversión texto a voz capaz de utilizar etiquetas predeterminadas es el sintetizador de Loquendo, donde se encuentran etiquetas para énfasis, velocidad del habla, expresiones del habla (“¡Hola!” “¡De acuerdo!” “¡Fantástico!”), estilos (enfático, formal, informal), emociones (alegre, triste, enojado), intención (por ejemplo: confirmación o duda) y eventos paralingüísticos (por ejemplo: respiración, tos, risa) [Loquen]. Otros sistemas de conversión texto a voz que poseen también estas posibilidades son el Mary [Sch03] o el RealSpeak de Nuance [Nuance]. Sin embargo, tanto la enumeración de los diferentes casos como el diseño del corpus sería complejo. Además, el hecho de enumerar los casos posibles restringiría la flexibilidad del sistema para abordar nuevos elementos prosódicos no considerados, y en cierta manera estaríamos reduciendo la realidad al modelo, en lugar de ajustar el modelo a la realidad.

Otro posible enfoque sería el utilizando en la aportación para la transferencia de la prosodia realizada en la tesis de Iriondo [Iri08]. Allí se utilizaron muestras de habla emocionada en castellano de cuatro emociones que tienen una expresión más universal (miedo, rabia, tristeza y alegría), de las que se extrajeron los valores de ciertos parámetros prosódicos que permitirían modificar el sistema TTS en catalán para generar habla emocionada en esta lengua. La información prosódica asociada a cada segmento en catalán se calculó usando los valores prosódicos de los segmentos de las locuciones del castellano: la energía y la F_0 se asignaron mediante un alineamiento temporal de sus contornos, las duraciones de las pausas se copiaron directamente, y la duración de las frases se ajustó globalmente; aplicando modificaciones a los segmentos proporcionalmente para cada emoción. Este enfoque asume que la transferencia de la prosodia puede ser directa. Sin embargo, es necesario realizar estudios en los aspectos prosódicos correspondientes entre dos idiomas para asegurar que esta asunción es correcta.

En esta tesis se pretende estudiar una alternativa consistente en una transferencia de

prosodia implícita. A través de un corpus paralelo se pretende encontrar relaciones entre lo que ocurre en la prosodia de un idioma a la vista de lo que ocurre en el otro idioma.

Para estudiar esta hipótesis se grabarán corpus usando hablantes bilingües, intentando preservar una coherencia en las diferentes decisiones de la realización prosódica en los distintos párrafos en ambos idiomas. Este enfoque no constituye una solución final, ya que un sistema de traducción no dispondrá de corpus para cada hablante, ni menos aún de corpus bilingüe. Todo este análisis se realizará para estudiar la viabilidad de la hipótesis con vistas a la generalización del caso dependiente del hablante a uno que sea independiente del hablante. En la Sección 5.3 se presentarán los algoritmos propuestos para la transferencia de la entonación en la traducción voz a voz.

Además de explorar esta hipótesis para la generación del contorno entonativo, también se lo hará para la duración y el ritmo. Es de interés para muchos sistemas de traducción automática voz a voz lograr un sincronismo entre la voz traducida y el video donde se ven tanto los labios del orador traducido como sus gestos (sin llegar a ser un doblaje), para evitar además el desfasaje entre los contenidos de los dos canales de información. De esta manera, estos documentos audiovisuales ganan en expresividad a través de la información complementaria expresada por el orador a través del lenguaje gestual, el ritmo y otros elementos prosódicos y visuales adicionales. Los enfoques propuestos al respecto se detallan en la Sección 5.4.

Finalmente, en la Sección 5.5 se realizarán aportes para la transferencia de pausas entre idiomas. Conjuntamente con la sincronización provista por la adaptación del ritmo, las pausas contribuyen tanto a segmentar el discurso en unidades más pequeñas (función delimitativa), como también a lograr la sincronización entre audio y video a través de la manipulación de la duración de las pausas

5.3. Generación de la entonación utilizando la información de la fuente

Como hemos comentado, las limitaciones de los ordenadores de hoy en día en lo referido a su desconocimiento sobre el mundo y la ausencia de ciertos rasgos de la psiquis humana (emociones, intención, complejos), así como también su incapacidad para procesar el lenguaje natural sin restricciones en la gramática (tal como ocurre, por ejemplo, en los sistemas de diálogo), contribuye a fijar un tope a la calidad alcanzable por los modelos de entonación que fueron tratados en las secciones 2.2 y 3.1.

Uno de los objetivos de esta tesis es la utilización de la información provista por la entonación del idioma origen para mejorar la naturalidad y expresividad de la entonación del idioma destino, reduciendo la necesidad de comprensión del lenguaje natural. Es más, si hay varias curvas entonativas adecuadas en cuanto a naturalidad y expresividad para un enunciado, queremos acercarnos a la que refleje mejor aquella utilizada por el locutor en el idioma origen.

El método propuesto consiste en la anotación del contorno origen utilizando información acústica y prescindiendo del contenido textual en la lengua origen. Debido a que la

información textual ya se encuentra en el texto traducido, no es necesaria la utilización del texto origen.

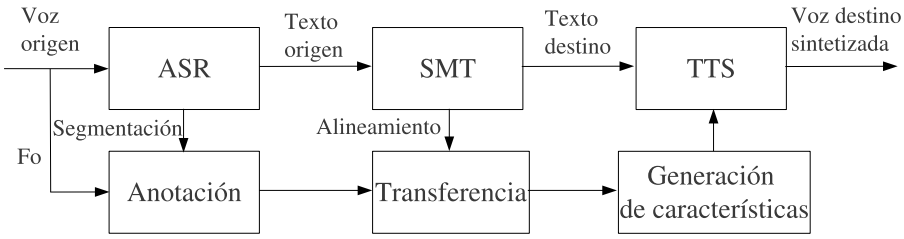


Figura 5.2: Esquema de generación de la entonación utilizando la voz fuente.

La Figura 5.2 es la particularización del esquema de transferencia prosódica mostrado en la Figura 5.1, para el caso de la entonación. En este caso, la información que se utiliza del habla fuente es su contorno entonativo. El sistema de reconocimiento del habla, mediante la segmentación temporal de las palabras reconocidas, permite asignar los contornos a cada palabra o la unidad entonativa que se utilice (por ejemplo grupo acentual, grupo entonativo, etc.) El sistema de traducción estadístico (SMT) puede producir no sólo la frase traducida, sino el alineamiento de las palabras entre ambos idiomas.

En este esquema se tiene el contorno de entonación de la frase fuente y se pretende generar un contorno para el texto traducido que utilice la información del texto traducido, pero también considere la información de la entonación en el idioma fuente. Para ello se podría utilizar características globales (tales como media, la varianza de f_0), pero es de interés en esta tesis el modelado de fenómenos locales, ya que estos son los que repercuten en mayor medida en las funciones de la prosodia explicadas al comienzo del capítulo.

Una posible implementación sería una proyección lineal entre contorno fuente y destino. Sin embargo, en la traducción el orden entra palabras no se mantiene, ni tampoco los contornos entonativos usados en las lenguas son los mismos. Por ello, es mejor intentar encontrar relaciones en la entonación de una palabra en una lengua en función de la entonación de la misma palabra en la otra lengua. Eso requiere trabajar palabra a palabra, y dicha información de alineamiento es proporcionada por el sistema de traducción estadístico.

El enfoque propuesto consiste en la transferencia de la entonación intentando encontrar relaciones entre los movimientos tonales a nivel de palabra. Es de esperar que un par de idiomas pueden compartir un conjunto de relaciones entre sus repositorios de movimientos tonales. Ciertos conceptos que se expresan con determinados contornos de entonación en un idioma son producidos en el otro idioma con un conjunto diferente de movimientos. Resultaría interesante encontrar tales relaciones entre idiomas para ayudar en la predicción de la entonación del idioma destino usando los movimientos tonales del idioma fuente. Además, es de esperar que aquellos idiomas más cercanos (por ejemplo: español y catalán) compartan más relaciones que aquellos idiomas que son más diferentes (por ejemplo: español e inglés). De este enfoque surgen una serie de desafíos que deben ser considerados para una apropiada implementación en un sistema de traducción voz a voz.

Con el objeto de validar la hipótesis que plantea la posibilidad de mejorar la expre-

sividad de la entonación del idioma destino usando el contorno de frecuencia fundamental del idioma origen, se decidió el diseño de corpus paralelos bilingües: catalán→español e inglés→español (Sección 5.3.1).

Otro elemento importante a considerar es el proceso de transferencia de la anotación del contorno del idioma origen para reflejar su influencia en la forma del contorno del idioma destino. Para ello la metodología propuesta usará la información de alineamiento entre los idiomas provista por el sistema de traducción automática (Sección 5.3.2).

La anotación del contorno origen debe ser automática y completamente acústica, considerando aquellos eventos en la entonación que serán de importancia para la generación de la entonación del idioma destino. Este tema será tratado en las secciones 5.3.3 y 5.3.4.

Finalmente, se realizó una validación experimental de la propuesta usando ambos corpus paralelos, para estudiar no solo su validez sino también la influencia ejercida por los orígenes de los idiomas en el rendimiento alcanzado (Sección 5.3.5).

5.3.1. Corpus orales para la investigación en generación de prosodia en traducción

En la actualidad los corpus orales bilingües son escasos, mientras que existe una gran disponibilidad de corpus escritos bilingües. Con el objeto de realizar experimentos sobre transferencia de prosodia se decidió la grabación de dos conjuntos de datos bilingües: inglés→español y catalán→español.

A diferencia de los que podría encontrarse en los intérpretes del Parlamento Europeo, una característica distintiva del corpus grabado es una mayor expresividad en cada idioma. En general, el traductor (o intérprete) no imprime en algunas ocasiones expresividad a su locución, con el objeto de que la proporcione la voz en segundo plano en el idioma original. Esta es una práctica muy utilizada, y es observada en parlamentos y foros internacionales.

La elección de los idiomas del corpus tiene su motivación en el estudio de la influencia de la cercanía entre idiomas en los resultados experimentales de nuestra propuesta. Es posible que idiomas de origen latino (español y catalán) posean recursos entonativos más similares que aquellos de origen distinto (español e inglés).

La investigación en esta tesis se limitó a sistemas monohablante, es decir, que tanto el hablante en un idioma como en el otro es una persona biligüe. En el futuro se explorarán mecanismos para su utilización en sistemas independientes del hablante.

Datos inglés→español

Para los experimentos inglés→español se diseñó un corpus bilingüe de 220 párrafos de textos correspondientes a párrafos parlamentarios paralelos. Cuatro hablantes bilingües (dos hombres y dos mujeres), pertenecientes a familias bilingües, grabaron los párrafos que corresponden a aproximadamente treinta minutos en cada idioma. El estilo de habla es parlamentario, usándose voces de parlamentarios reales para indicar al hablante el estilo deseado para los párrafos. Los párrafos provenían de sesiones distintas buscando

estilos variados. Estas voces grabadas fueron finalmente uno de los recursos generados en el proyecto TC-STAR.

El estudio de datos bilingües exige una consistencia en los datos. A los hablantes bilingües se les requirió grabar cada párrafo en un idioma, e inmediatamente el correspondiente en el otro idioma, manteniendo el estilo y la consistencia del discurso entre idiomas.

La anotación prosódica de los párrafos consta de dos niveles de junturas terminales y dos niveles de énfasis de palabra. A continuación se adjunta como ejemplo un párrafo bilingüe:

> Inglés: "The commission's mandate is for the running of the European Union and yet there seem to be some in this house who would have it closed down for the summer. Do they not envisage that the settlement of the European Union's budget for the next seven years or the question of Turkish membership might make some demands on the commission's attention between now and November?"

> Español: "El mandato de la comisión es que gestione la Unión Europea, pero parece que en esta cámara hay quien quisiera cerrarla durante el verano. ¿Acaso no tienen previsto que la concreción del presupuesto de la Unión Europea para los próximos siete años o la cuestión del ingreso de Turquía pueda reclamar la atención de comisión de aquí a noviembre?"

Datos catalán→español

Los datos de catalán→español son un recurso generado para el proyecto LC-STAR. Un hablante bilingüe grabó 40 minutos de diálogos en el dominio de agencia de viaje. Los datos son paralelos y el estilo de habla es natural.

Las doscientas oraciones que componen el corpus no tienen información de puntuación, siendo los puntos finales los únicos disponibles. Las junturas terminales fueron etiquetadas manualmente usando un solo nivel: grupos entonativos.

Aquí también se prestó especial atención a la consistencia entre idiomas, tal como se mencionó en los datos de inglés→español,

> Catalán: "Ha de donar-me el número de la targeta de crèdit i la data en què li caduca la targeta de crèdit."

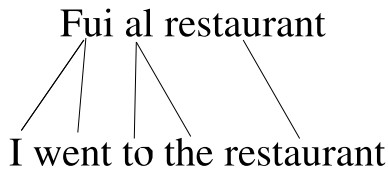
> Español: "Tiene que darme el número de la tarjeta de crédito y la fecha en que le caduca la tarjeta de crédito."

5.3.2. Transferencia de información del contorno origen para generar el contorno destino

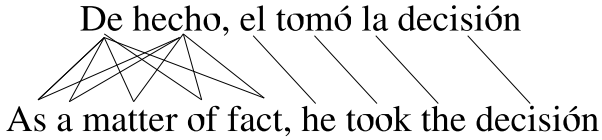
La información sobre alineamiento proporcionada por el módulo de traducción automática es de gran importancia para el uso de la información del contorno origen para generar la entonación del idioma destino.

El alineamiento proporciona vínculos entre palabras, y cada una de ellas puede poseer desde ninguno hasta múltiples vínculos con palabras del otro idioma. Los casos que pueden surgir se enumeran a continuación:

- *Una a muchas.* Las palabras en el idioma origen se puede traducir a una o más palabras en el idioma destino debido a razones léxicas, gramaticales, sintácticas o semánticas. Por ejemplo, la oración "Fui al restaurante" que se traduce al inglés como "I (NULL) went (fui) to the (al) restaurant (restaurante)", "al" se alinea con dos palabras debido a que en español "a el" se contrae a "al".



- *Muchas a muchas.* Algunas construcciones del idioma origen deben estar alineadas con su contraparte en el otro idioma para preservar el significado. El significado individual de las palabras difiere de su uso conjunto. Por ejemplo, la oración "De hecho, él tomó la decisión." se traduce como "As a matter of fact (De hecho), he (él) took (tomó) the (la) decision (decisión).". Aquí, "As a matter of fact" se traduce como "De hecho" en español. No es posible alinear estas construcciones en sus palabras constituyentes sin perder el significado global.



- *Sin alineamiento.* En algunas situaciones una palabra en un idioma no puede ser alineada con ninguna del otro idioma. Estas palabras son solamente usadas en uno de los idiomas, y no tienen una contraparte en el otro idioma.

Mediante operaciones de conjuntos es posible obtener una simplificación de los vínculos para tener un alineamiento uno a uno a nivel de palabras, tal como ocurre con el uso del operador intersección (para más detalles se puede leer Och et al. [Och00]). A continuación se muestra una porción de textos alineados usando el operador de intersección (se incluye

un número indicando el orden de las palabras en el texto del idioma origen para facilitar la lectura):

(1)señor (2)presidente (3)en (4)nombre (5)del (6)grupo (7)del (8)partido (9)europeo 10)de (11)los (12)liberales (13)democratas (14)y (15)reformistas

(1)mr (2)president (3)on (4)behalf (5)of ()the (9)european (12)liberal (13)democrat (14)and (15)reform (8)group

El uso del operador intersección y el alineamiento de palabras resulta de gran utilidad, ya que permite analizar los eventos entonativos a nivel de palabra y estudiar su relación con la forma de la entonación de la palabra en el otro idioma.

Para el análisis de la relación entre contornos entonativos de dos idiomas es necesario elegir una unidad prosódica que delimite un entorno de tiempo para su estudio. La información sobre alineamiento será usada para tener en cuenta las diferencias en el ordenamiento de las palabras entre idiomas y poder estudiar los movimientos tonales correspondientes. Dentro de las unidades que pueden ser seleccionadas se encuentran la sílaba, el grupo acentual y la frase entonativa. La elección de la unidad prosódica debe tener en cuenta diferentes factores que se relacionan con la prosodia y con la tarea específica de la traducción voz a voz. La unidad prosódica elegida debe ofrecer una buena cobertura tanto en términos de expresividad como en su facilidad para analizar las relaciones entre los idiomas teniendo en cuenta el posible orden diferente de las palabras.

La sílaba es una unidad prosódica que ofrece un poder de análisis local con alta resolución. Sin embargo, no es adecuada para nuestros propósitos debido a que no es posible encontrar una correspondencia entre sílabas de diferentes idiomas a través del alineamiento. El alineamiento solamente proporciona vínculos entre palabras (por ejemplo: casa $\rightarrow$ house) o grupos de palabras (por ejemplo: sin embargo $\rightarrow$ however).

La frase entonativa no puede ser tenida en cuenta porque es una unidad prosódica muy amplia que puede contener muchos movimientos tonales causados por diferentes razones. Por ejemplo, una frase entonativa puede contener una palabra enfatizada y un movimiento tonal ocasionado por el estado de ánimo en otro grupo de palabras.

En esta tesis se optó por el grupo acentual como la unidad prosódica elegida. El mismo permite modelar habla expresiva si la función de aproximación posee la flexibilidad necesaria para ajustarse a los movimientos tonales del grupo acentual. Además, es posible obtener relaciones entre los grupos acentuales de los idiomas si usamos la palabra acentuada como referencia, tal como se definió para el español en la Sección 2.2.1. Un ejemplo de alineamiento usando el grupo acentual se muestra en la Figura 5.3.

5.3.3. Sistemas de anotación simbólica de la entonación

El proceso de generación de la entonación usando información sobre los contornos de frecuencia fundamental del idioma fuente sugiere la utilización de una codificación de los movimientos tonales, representando mediante símbolos ciertas configuraciones de

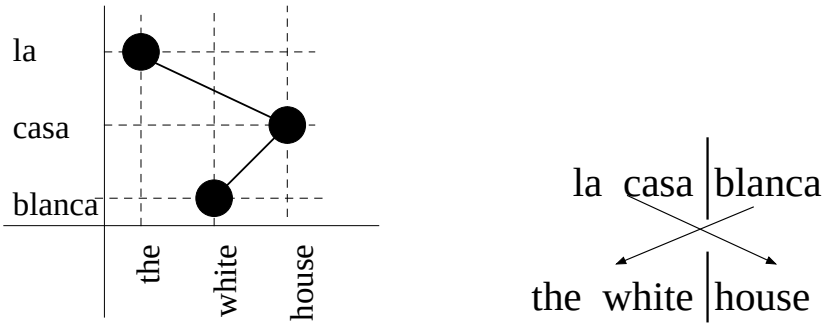


Figura 5.3: Alineamiento usando grupos acentuales.

la curva entonativa. De esta manera, esta información podría ser transferida al idioma destino, constituyendo una característica adicional para la generación de la entonación en el conversor texto a voz.

En general, se puede afirmar que esta codificación se puede realizar usando un conjunto predefinido de tonos (y símbolos que los representan), o bien definir automáticamente un conjunto limitado de símbolos mediante técnicas de agrupamiento automático.

En la literatura existen diversos ejemplos de sistemas de codificación de movimientos tonales usando un conjunto predefinido de símbolos. Uno de ellos es ToBI [Sil92], que fue utilizado en sus comienzos para codificar la entonación del inglés americano. ToBI especifica seis tipos de tonos para el inglés americano: H*, L*, L*+H, L*+H, L+H* y !H*. Los niveles H(alto) y L(bajo) indican un punto relativamente alto o bajo en el rango del hablante, y pueden concatenarse para señalar una combinación usando el símbolo +. El símbolo * indica un alineamiento directo con la sílaba acentuada. Por otra parte, también existe una codificación para las junturas terminales, que difieren dependiendo del tipo de movimiento tonal y la intensidad percibida de la juntura: L-, H-, L-L %, L-H %, H-H %, H-L % y %H. Por ejemplo, la tendencia de la sílaba final de las oraciones declarativas se puede señalar a través de un movimiento tonal hacia el punto inferior del rango del hablante: L-L %.

Es importante establecer ciertos requisitos de precisión en la transcripción de los movimientos tonales para su utilización en la caracterización de contornos, y en esta dirección existen experimentos demostrando un alto grado de acuerdo entre transcriptores diferentes de ToBI [Pit94]. Sin embargo, otros estudios sobre ToBI contradicen los resultados de Pitrelli, y por el contrario señalan que la concordancia entre transcriptores es relativamente baja [Wig02]. Un estudio entre seis transcriptores entrenados de manera uniforme y con acceso a contornos con alineado temporal, espectrogramas y señales de audio demostró que las etiquetas coincidían en menos del 50 % para seis de las ocho etiquetas bajo estudio [Syr00]. Estos estudios señalan que la precisión del etiquetado en ToBI es solamente alta para un subconjunto de etiquetas. Otro aspecto negativo sobre la utilización de ToBI se relaciona con los tiempos de etiquetado, que pueden llegar a alcanzar de 100 a 200 veces la duración del audio analizado [Syr01].

No existen en la actualidad métodos automáticos fiables para analizar el contorno de entonación de un idioma usando ToBI. Teniendo en cuenta estos inconvenientes es poco

viable la utilización de ToBI como sistema de codificación de los contornos del idioma fuente. Los datos de entrenamiento para un sistema automático de transcripción de ToBI necesario en una plataforma de traducción voz a voz tendrían muchas inconsistencias debido al poco grado de acuerdo entre transcriptores humanos.

Otro sistema de codificación descrito en la literatura es INTSINT [Hir94]. Esta codificación señala los eventos significantes de la curva tonal usando un conjunto limitado de símbolos para señalar tonos absolutos (T, M y B) y relativos (H, L, S, U y D). Los tonos absolutos en INTSINT se definen de acuerdo al rango tonal del hablante, mientras que los relativos se anotan con respecto a la altura tonal de los puntos adyacentes. En su conjunto permiten hacer una descripción detallada del contorno de frecuencia fundamental a través del análisis automático de la entonación [Hir00] usando una herramienta de estilización de contornos: MOMEL [Hir93]. Uno de los aspectos remarcables de INTSINT es que la transcripción conserva los valores numéricos de los eventos tonales. Por lo tanto, es posible representar la curva tanto en forma cualitativa (como en el caso de ToBI) como cuantitativa (parametrizada). Las correlaciones lingüístico/funcionales de estos eventos pueden vincularse con un análisis de las propiedades pragmáticas, semánticas y sintácticas de la oración.

Sin embargo, la utilización de un sistema simbólico tal como INTSINT podría resultar insuficiente. El conjunto de símbolos es muy limitado (representación cualitativa), y la capacidad para modelar la diversidad se sustentaría en la representación numérica (representación cuantitativa). Esta última no constituye un conjunto discreto, y por lo tanto, no resulta viable para su utilización como codificación simbólica. Sin embargo, este enfoque podría ser útil en un sistema que combine codificación simbólica y técnicas de regresión.

En esta tesis se propone la utilización de un algoritmo automático de agrupamiento para encontrar estas correspondencias. El mismo se basa en el total desconocimiento acerca de las relaciones entre contornos de entonación de los idiomas. De esta manera los patrones de entonación serán obtenidos con una etiqueta abstracta, sin significado aparente. Posteriores estudios, fuera del alcance de esta tesis, deberán analizar esa información para encontrar el significado y la relación de tales patrones.

Es necesario aclarar que muchos contornos de entonación del idioma fuente pueden tener una correspondencia con muchos contornos de entonación en el idioma destino, un tema que ya fue analizado en el capítulo sobre generación de prosodia. En esta tesis nos enfocaremos en una correspondencia uno a uno. A pesar de que este enfoque es limitado, es un punto de partida para futuras investigaciones.

En la siguiente sección explicaremos el entrenamiento de los algoritmos de anotación automática. En la primera fase se encuentran patrones (contornos de entonación típicos) que relacionan los movimientos tonales de los idiomas en los datos de entrenamiento. Estos patrones luego son usados para etiquetar los contornos de entonación en el sistema completo de generación de prosodia.

5.3.4. Anotación de la entonación del hablante fuente

El objetivo del algoritmo de anotación consiste en agrupar aquellos movimientos tonales del idioma origen que consistentemente tienen una correspondencia con movimientos del idioma destino. Tales movimientos se consideran patrones que se repiten en la base de datos (clases de movimientos tonales) y se pueden usar para codificar la entonación de la entrada. La codificación es útil debido a que puede ser usada como característica de entrada adicional para el módulo de generación de la entonación. Es esperable que proporcione importante información semántica que mejore la naturalidad y la expresividad de la entonación. El algoritmo de agrupación debe tener en cuenta que algunos movimientos del idioma fuente no tienen una correspondencia en el idioma destino. En tales casos no es posible asignar a los movimientos tonales clase alguna. Como consecuencia de esto, los mismos corresponderán a la clase SINCLASE.

Por otra parte, el alineamiento automático entre idiomas puede provocar que una palabra o un grupo de ellas no se encuentren alineadas. En este caso el algoritmo no podrá encontrar relaciones para tales movimientos tonales y corresponderán también a la clase SINCLASE.

La figura 5.4 muestra una hipotética frase de entrada bilingüe con el alineamiento definido por el sistema de traducción. El objetivo del algoritmo de anotación es detectar que hay una correspondencia, por ejemplo, entre el grupo acentual 2 del idioma origen y el grupo acentual 4 del idioma destino. Por ello, ambos se etiquetarían con las clases 2 y 2*. Hay otros grupos acentuales, como el grupo acentual 3 del origen alineado con el grupo acentual 2 del destino, donde se observa que el movimiento no debe ser consistente en el corpus, y por eso se les ha asignado la etiqueta SINCLASE. Y finalmente, el grupo acentual 4 del idioma origen ni siquiera esta alineado con un grupo acentual del idioma destino, y por ello también se le ha asignado la etiqueta SINCLASE.

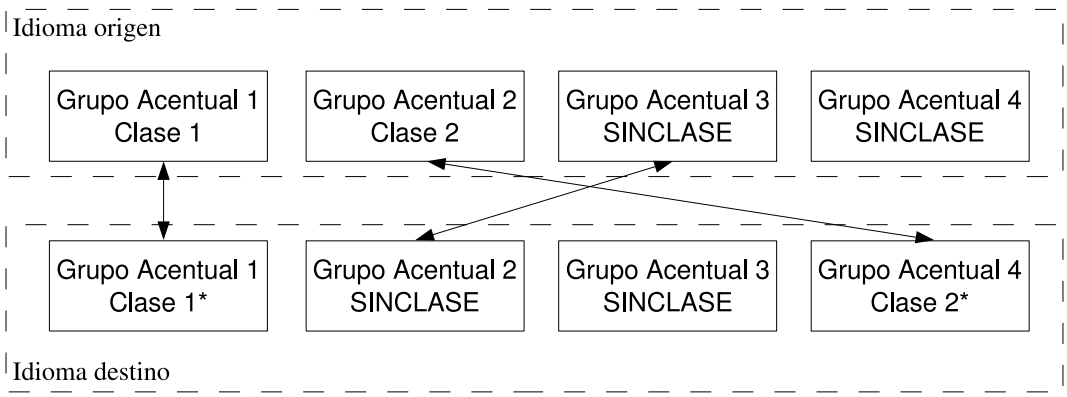


Figura 5.4: Ejemplo de alineamiento de grupos acentuales y asignación de clases.

El algoritmo de agrupamiento que proponemos encuentra un número arbitrario de clases (o patrones) en forma iterativa, partiendo de umbrales para la comparación poco estrictos, que luego se irán ajustando para lograr relaciones entre clases realmente significativas.

En este apartado se explicarán dos algoritmos. El primer algoritmo (Sección 5.3.4.1)

enfatiza la búsqueda de relaciones entre idiomas usando una distancia que incluye a los contornos de entonación origen y destino. Este algoritmo tiene un buen rendimiento para los datos de entrenamiento pero presenta problemas de generalización debido a una tendencia a sobrestimar los datos de entrenamiento. El segundo algoritmo (Sección 5.3.4.2) propuesto soluciona este problema a través de una búsqueda de las relaciones entre idiomas minimizando el error de aproximación en la clases del idioma destino para los datos de entrenamiento y usando una distancia que solo incluye al idioma origen.

5.3.4.1. Anotación por similitud de contornos tanto en el idioma fuente como en el destino.

El algoritmo de agrupamiento intenta encontrar movimientos tonales similares tanto en el idioma fuente como en el idioma destino usando el error cuadrático medio como medida objetiva. Aquel patrón de clase que mejor aproxime a cada movimiento tonal en RMSE para los datos de entrenamiento, será la clase correspondiente de dicho movimiento.

El RMSE se calcula conjuntamente entre los contornos origen y destino, y los patrones de clase correspondientes. Esta medida conjunta se usa para enfatizar el hecho que el algoritmo buscará correspondencias entre idiomas. El objetivo es obtener un conjunto de clases que tendrán una buena cobertura (muchos grupos acentuales de los datos de entrenamiento pertenecerán a la clase) y un RMSE bajo entre los patrones de las clases de ambos idiomas y los contornos que pertenecen a las mismas.

El primer algoritmo propuesto intenta encontrar relaciones entre los movimientos tonales de ambos idiomas en forma iterativa. El mismo consiste en dos ciclos de mejora. En el primero de ellos (ciclo B) se intenta optimizar la calidad de las clases obtenidas, mientras que en el segundo (ciclo A) se incorporan nuevas clases y se eliminan aquellas que son poco representativas, tal como se observa en la Figura 5.5. El algoritmo es similar al LBG utilizado en el diseño de codebooks [Lin80].

Un umbral en la distancia euclídea entre el contorno y el patrón de la fuente contribuye a eliminar contornos que no deberían pertenecer a una clase. De esta manera, estos son asignados a la clase SINCLASE. El umbral tiene un valor inicial de cuatro semitonos y decrece exponencialmente hasta alcanzar dos semitonos luego de veinte iteraciones. Los contornos eliminados puede ser ubicados en otras clases en futuras iteraciones del ciclo A.

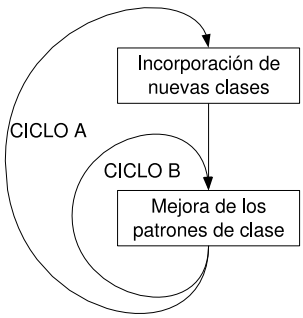


Figura 5.5: Ciclo de mejora continua de clases.

En cada iteración del ciclo B se recalculan los contornos de los patrones de clase con el fin de minimizar el error de aproximación con respecto a todos los contornos que pertenecen a dicha clase. Esta redefinición de los patrones provoca una posible reasignación de los contornos a otras clases, y se itera en el ciclo B hasta que se produzca una convergencia, que equivale a una disminución del número de reasignaciones por debajo de un umbral. Luego, en el ciclo A se añaden nuevas clases, y el proceso del ciclo B se repite.

Con el objetivo de aclarar en forma detallada el funcionamiento del algoritmo, a continuación se detallan los pasos del mismo:

1. *Inicialización.* La inicialización consiste en asignar clases aleatorias a cada grupo acentual en el idioma origen y etiquetar los grupos acentuales del idioma destino con la clase correspondiente de acuerdo a la información de alineamiento. Por ejemplo, la **clase 1** del idioma origen se vincula con la **clase 1*** del idioma destino, la **clase 2** se vincula con la **clase 2***, y asi sucesivamente. En la inicialización, los grupos acentuales que no tienen vínculos entre idiomas son asignados a la clase SINCLASE.

2. *Patrones óptimos.* Se ha representado los contornos de cada clase mediante polinomios de Bézier de grupos acentuales, estimados siguiendo el método de optimización explicado en la Sección 3.1.3 sobre modelado de la entonación basado en Bézier no-superposicional entrenado usando el enfoque JEMA. Este algoritmo permite obtener los patrones que aproximan de forma óptima todos los contornos que pertenecen a una clase dada. Es necesario aclarar que el patrón óptimo puede ser diferente de muchos de los contornos que pertenecen a la clase, debido a que en las primeras iteraciones hay una gran dispersión en los contornos asignados a cada clase.

3. *Clasificación de los movimientos tonales.* Aquí los patrones obtenidos en el paso previo se usan para clasificar los movimientos tonales de todos los contornos. Entonces, muchos movimientos pueden cambiar su clase asignada debido a que otro patrón lo aproxima en forma más precisa teniendo en cuenta el error conjunto en los idiomas origen y destino. De esta manera el algoritmo de clasificación encuentra movimientos tonales que tienen relación entre los idiomas. El error se mide usando una distancia euclídea con la frecuencia en escala logarítmica.

4. *Convergencia del algoritmo.* En este paso se calcula el número de movimientos tonales que cambiaron su clase. Esto es una medida de la convergencia del algoritmo. Si el porcentaje de cambios sobre el número total de movimientos tonales es inferior a un umbral (en este trabajo 0,1 %), consideramos que el algoritmo convergió y se sale del ciclo B. Los movimientos tonales que pertenecen a una clase que tiene menos de cuarenta elementos también se consideran que pertenencen a la clase SINCLASE. Una clase con pocos elementos es poco representativa debido a que no hay información suficiente en los datos de entrenamiento para justificar su existencia.

5. *Agregado de nuevas clases* En este momento se agregarían nuevas clases para comenzar otro ciclo A. Los movimientos tonales que fueron asignados a la clase SINCLASE se asignan aleatoriamente a alguna de las dos nuevas clases agregadas. De esta manera, aquellos movimientos que no pertenecen a ninguna clase pueden generar clases nuevas.

6. *Condición de parada.* El algoritmo propuesto se considera que ha convergido al número necesario de clases cuando el RMSE de los contornos del lenguaje destino en los datos de entrenamiento no se reduce en un número preestablecido de iteraciones del ciclo A (en nuestros experimentos este número es de diez).

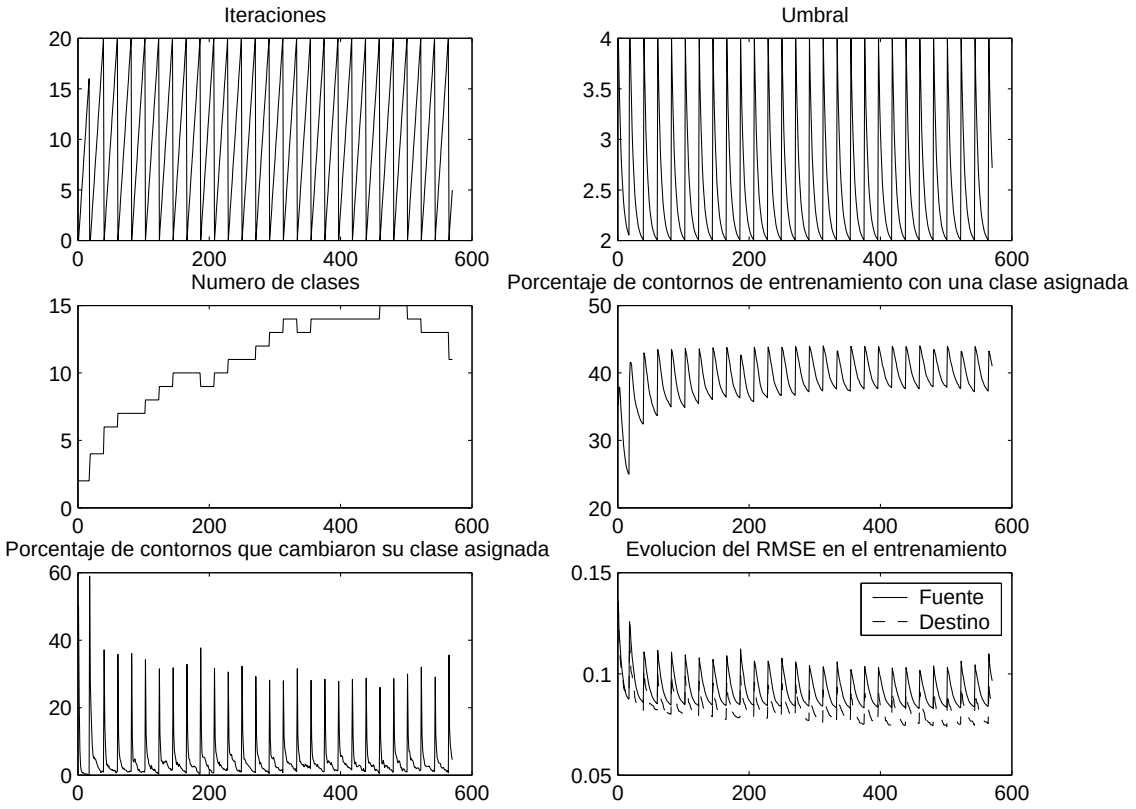


Figura 5.6: Evolución de los parámetros de entrenamiento durante el agrupamiento.

En la Figura 5.6 mostramos la evolución de varios parámetros del algoritmo de agrupamiento en un experimento usando el inglés como idioma origen y el español como idioma destino, usando 8000 grupos acentuales.

El primer gráfico muestra la evolución del número de iteraciones del ciclo B. Luego de veinte iteraciones, o si el porcentaje de cambios (mostrado en el tercer gráfico de la izquierda) está debajo de 0,1 %, dos nuevas clases se agregan (ciclo A). Las rampas que se observan tienen su origen en el valor del contador del número de iteraciones, que una vez inicializado o reinicializado vuelve a tomar el valor 0.

En el segundo gráfico de la izquierda se puede ver el aumento del número de clases al comienzo de cada iteración. Es de destacar que cerca de la iteración 200 el número de clases decrece momentáneamente. Este es el efecto de la desaparición de clases debido a su baja representatividad.

En el primer gráfico de la derecha se puede ver el decrecimiento exponencial del umbral de RMSE para establecer la pertenencia de un movimiento tonal a una clase dada, cuya

evolución se eligió arbitrariamente para que comience con un valor elevado (baja selectividad) y para terminar con un valor bajo (mejor representatividad). La evolución de este umbral tiene un impacto directo en el número de movimientos asignados a cada clase, y en consecuencia, en el porcentaje de contornos con una clase asignada (segundo gráfico de la derecha).

La dinámica de cambio de clase debido a la mejor representatividad de un patrón que otro se muestra en la última gráfica de la izquierda. En cada ciclo B, el número de clases que cambian al principio es elevado, disminuyendo continuamente hasta alcanzar las 20 iteraciones (o el umbral prefijado de 0,1 %).

En el último gráfico de la derecha se observa el RMSE entre los patrones obtenidos y los contornos pertenecientes a cada clase, tanto para el idioma fuente como para el destino. No se han descartado los contornos asignados a la clase SINCLASE, y por ello el RMSE aumenta con el agregado de clases.

El RMSE de cada última iteración del ciclo B para el hablante destino es utilizado para establecer la condición de parada. En este ejemplo, dicho RMSE no se reduce a partir de la iteración 400 durante diez iteraciones del ciclo A, provocando que se cumpla la condición de parada.

5.3.4.2. Anotación por similitud de contornos en el idioma fuente.

El segundo algoritmo propuesto tiene los mismos objetivos que el primero en lo referente a encontrar patrones entonativos relacionados entre los idiomas para mejorar la expresividad de la síntesis de voz. En el primer algoritmo se puso énfasis en la utilización de una distancia conjunta que involucre a los contornos de los idiomas fuente y destino con el objeto de encontrar vínculos entre movimientos tonales durante el ciclo B. Sin embargo, en el momento del uso de las clases en un sistema de traducción voz a voz, la única distancia que se calculará para encontrar la clase a la que corresponde cada contorno será la obtenida con el idioma fuente. Por tanto, no sería útil tener patrones distintos que se diferencien en el idioma destino, ya que no podríamos elegir en el momento de funcionamiento del sistema de traducción voz a voz cual es el que corresponde a una situación en particular.

Por lo tanto, en este segundo algoritmo se propone la utilización de una distancia que consista solamente del contorno del idioma fuente durante el ciclo B. De esta manera se orienta la búsqueda de relaciones sin usar diferentes distancias durante el entrenamiento y en el uso posterior de las clases en la traducción voz a voz.

Una consecuencia importante de este nuevo enfoque es que se evitará un fenómeno que puede llegar a suceder en el algoritmo propuesto en la sección anterior, que consiste en la aparición de clases debido a patrones diferentes del idioma destino que no tienen diferencias en los patrones de clase del idioma origen. Cuando ello ocurriese para un contorno que pertenezca a alguna de esas clases del idioma origen, se elegiría una clase del idioma destino en forma aleatoria, lo cual no siempre resultaría correcto.

De todas maneras, para lograr encontrar relaciones de patrones entre idiomas, se continuarán eliminando aquellas clases con un alto error de aproximación entre el contorno patrón de clase y los contornos que pertenecen a dicha clase, tanto para el idioma fuente

como para el destino. Un patrón de clase con un alto RMSE es poco representativo y es un indicador de la baja consistencia de esa clase.

A través de este proceso de eliminación de clases con un alto error de aproximación tanto para el idioma fuente como para el destino, se fortalece la búsqueda de aquellas clases que permitirán estimar la entonación del idioma destino basándose en los contornos del idioma fuente.

5.3.5. Validación experimental

5.3.5.1. Condiciones experimentales

Los algoritmos de extracción de patrones fueron evaluados a través de varios experimentos con diferentes condiciones. En todas las situaciones los algoritmos propuestos se comparan con un sistema base que no hace uso de la información adicional de la entonación del hablante origen. De esta manera se puede observar la ganancia obtenida con el agregado de la codificación de los movimientos tonales. Los datos experimentales se dividieron en diez partes para realizar *10-fold cross-validation*.

Tanto la transcripción ortográfica como la traducción son correctos, lo cual supone una tasa de error en el reconocimiento de 0 % y una traducción perfecta. El uso de esta situación ideal constituye una primera aproximación al problema, evitando posibles inconvenientes en el análisis debido al ruido introducido por errores en ASR y SMT. La información de alineamiento fue provista por GIZA++ [Och03], utilizando un corpus más grande e incluyendo la información que se quería alinear dentro del mismo. De esta manera se obtuvieron los vínculos entre palabras de los idiomas.

El modelo prosódico utilizado es el explicado en la Sección 3.1.3: modelo de entonación basado en Bézier no-superposicional entrenado usando el enfoque JEMA. Las características extraídas del texto traducido son las mismas a las utilizadas en la Sección 4.2.1: posición dentro del grupo entonativo, la posición de la sílaba acentuada, el número de sílabas y palabras que lo constituyen, e información sobre signos de puntuación en sus fronteras (en caso de que se encuentren disponibles).

Los dos primeros experimentos se realizaron con los datos inglés→español, con el objeto de estudiar el rendimiento de los dos algoritmos propuestos en dos idiomas de origen distinto.

Experimento 1

En este experimento se usaron los patrones de clase obtenidos usando el primer algoritmo de agrupación con información acerca de los signos de puntuación para el modelado de la entonación.

Los resultados de los experimentos usando el primer algoritmo propuesto con los datos correspondientes a inglés→español se muestran en la Figura 5.7.

La figura muestra el diagrama *boxplot* del RMSE para el algoritmo propuesto (P.)

y para el algoritmo base (B.) para los cuatro locutores utilizados en los experimentos. El algoritmo base no posee información de la codificación de la entonación del hablante origen. Los resultados se muestran tanto para los datos utilizados en definir la anotación y el entrenamiento del modelo prosódico (entrenamiento), como también para los datos reservados para la evaluación (evaluación).

En cada uno de los gráficos correspondientes a cada uno de los cuatro locutores se puede observar que el algoritmo propuesto (**P. entrenamiento**) presenta un mejor desempeño con los datos de entrenamiento que aquel modelo que no utiliza la información de la codificación de la entonación del hablante origen (**B. entrenamiento**). Esto muestra la mejora alcanzada con el uso de la nueva característica prosódica.

Sin embargo, el RMSE en los datos de evaluación no presentan diferencias entre las condiciones **P. evaluación** y **B. evaluación**. El primer algoritmo propuesto no tiene buenas propiedades de generalización, con una tendencia a sobrestimar el modelo a los datos de entrenamiento. Por lo tanto, no puede ser usado para los propósitos de la transferencia de prosodia.

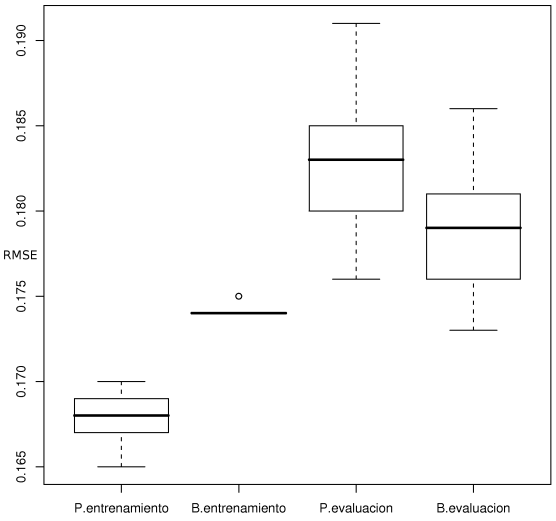
Experimento 2

En este experimento se usaron los patrones de clase obtenidos usando el segundo algoritmo de agrupación con información acerca de los signos de puntuación para el modelado de la entonación.

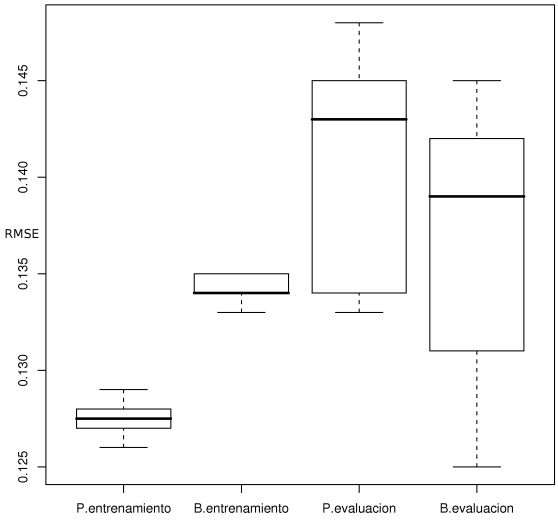
La Figura 5.8 muestra los resultados experimentales usando el segundo algoritmo propuesto. En este caso la ganancia obtenida para los datos de evaluación es ligeramente mejor que el primer algoritmo en lo relativo a la mediana, pero el tercer cuartil señala que la mejora no es estadísticamente significativa.

La ganancia es muy pequeña en casi todos los casos, lo que puede provocar que la mejora no sea perceptible por el oyente. Estos resultados son motivados por diversas razones:

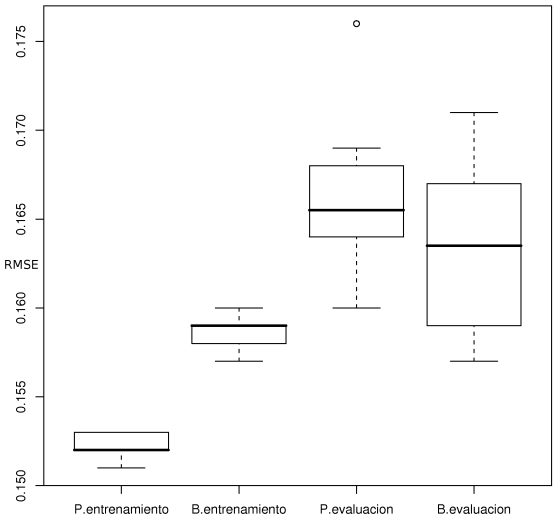
- *Idiomas pertenecientes a diferentes ramas.* El español y el inglés son idiomas que pertenecen a diferentes ramas. El español es un idioma latino mientras que el inglés es un idioma germánico. Esta diferencia se ve reflejada en todos los aspectos del idioma, y la prosodia no es una excepción. Algunos recursos prosódicos pueden estar presentes en un idioma y estar ausentes en el otro.
- *Múltiples contornos con el mismo significado.* En la Sección 3.1.1 se explicó que uno de los problemas del modelado de la entonación es que muchos contornos pueden ser usados por el hablante para expresar el mismo significado. En el caso de la transferencia de prosodia muchos contornos del idioma origen pueden estar relacionados con muchos del idioma destino. Este es un problema que debe ser enfrentado en futuras propuestas.
- *Alineamiento.* El orden de las palabras en español e inglés es diferente. Esto introduce algunos problemas para la transferencia de prosodia porque los movimientos tonales pueden diferir debido al ordenamiento de las palabras.



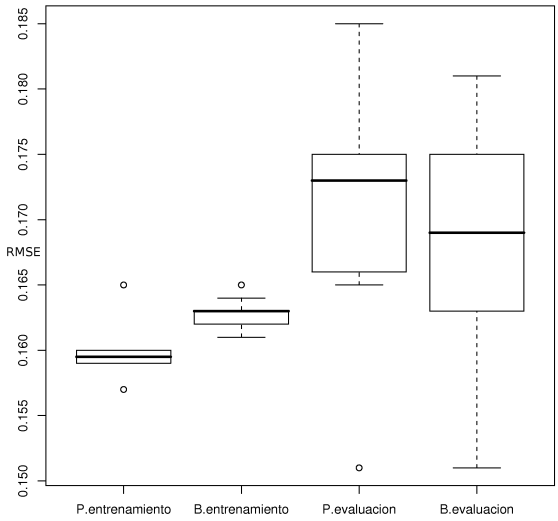
Mujer 1



Mujer 2



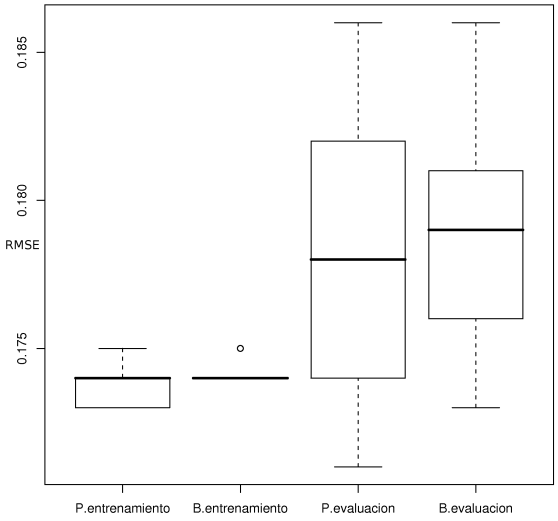
Hombre 1



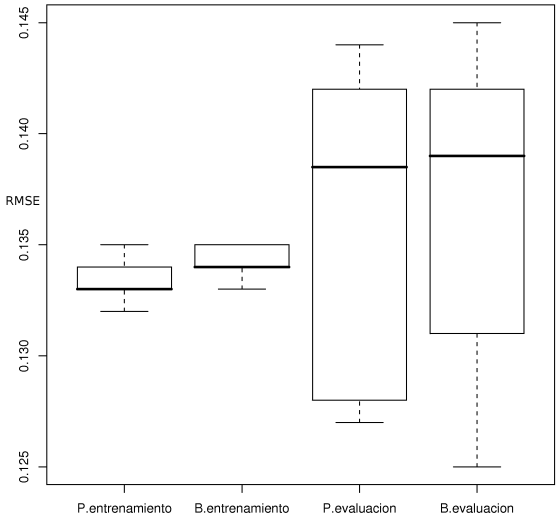
Hombre 2

Figura 5.7: Resultados experimentales usando el primer algoritmo propuesto en la dirección inglés → español.

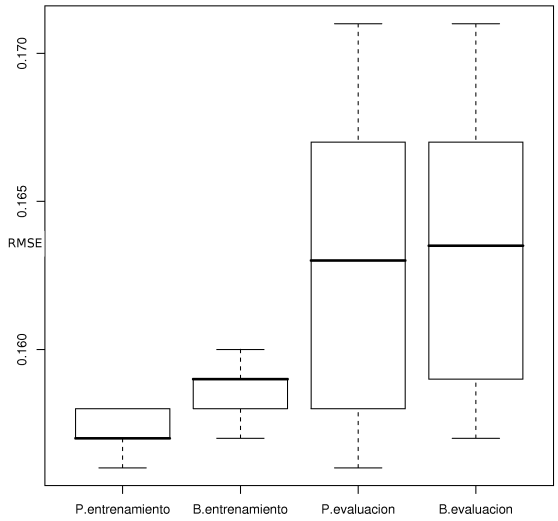
- *Alineamiento automático.* La información de alineamiento provista por GIZA cubre solamente el 72 % de las palabras del idioma origen. Este hecho indica que muchas relaciones entre los idiomas no están siendo estudiadas.
- *Componente de grupo entonativo.* La transferencia de entonación se realizó usando al



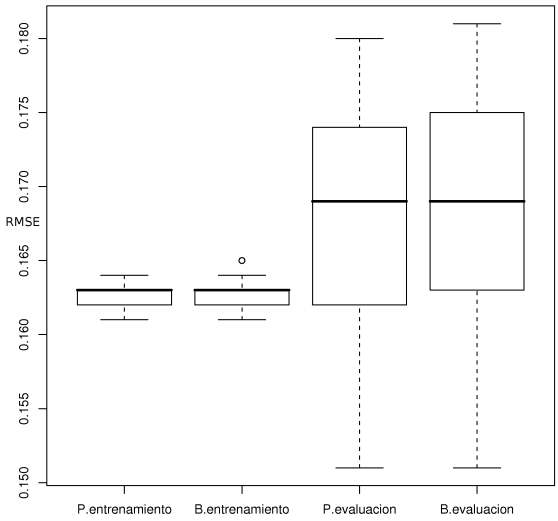
Mujer 1



Mujer 2



Hombre 1



Hombre 2

Figura 5.8: Resultados experimentales usando el segundo algoritmo propuesto en la dirección inglés → español. Los datos poseen todos los signos de puntuación.

grupo acentual como unidad prosódica. Sin embargo, el grupo acentual está afectado por la influencia de la componente de grupo entonativo. Por lo tanto, será necesario desarrollar algún procedimiento para separar este efecto y analizar el grupo acentual sin la perturbación de otras componentes.

- *Consistencia.* A pesar de que los datos bilingües fueron grabados prestando cuidadosa atención a la consistencia prosódica, es posible que algunas de ellas estén presentes debido a que algunos párrafos eran largos. Es difícil para el hablante recordar el estilo usado en cada palabra en tales situaciones.

Experimentos sobre la importancia del origen de los idiomas involucrados

El catalán y el español son idiomas cercanos y es esperable que la transferencia de prosodia sea una tarea más fácil. En la Figura 5.9 se muestran los resultados experimentales correspondientes al uso del segundo algoritmo propuesto. El primer algoritmo no ha sido evaluado en este caso debido al problema de generalización descrito en los experimentos anteriores.

La gráfica indica que tanto para los datos de entrenamiento como para los de evaluación el algoritmo propuesto permite obtener un menor error cuadrático medio en la estimación de los contornos de entonación del hablante destino.

La ubicación del tercer cuartil para el algoritmo propuesto y los datos de evaluación por debajo del primer cuartil del algoritmo base es un indicador de una mejora significativa. Esta diferencia debería ser perceptible por el oyente, porque tal mejora con respecto a los resultados del sistema base son estadísticamente significativas: $p < 0{,}01\,\%$ para los datos de entrenamiento y $p < 2\,\%$ para los datos de evaluación.

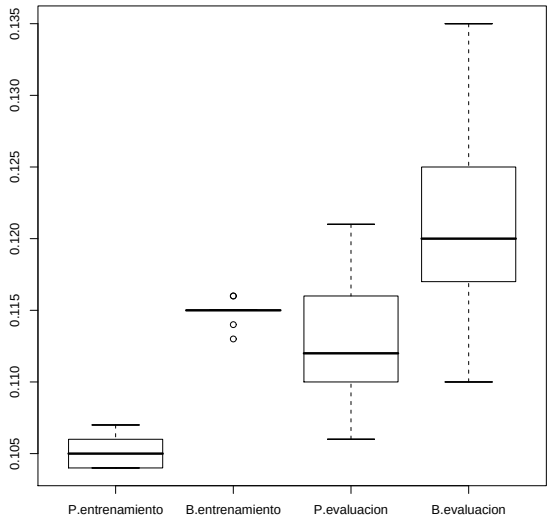


Figura 5.9: RMSE del logaritmo de la frecuencia fundamental usando el segundo algoritmo propuesto en la dirección catalán → español. Los datos solamente poseen puntos finales.

Las razones para la mejora en el rendimiento usando estos datos son:

- *Idiomas pertenecientes a la misma rama.* Tanto el catalán como el español son id-

iomas latinos.

- *Alineamiento.* El ordenamiento de las palabras en español y en catalán es prácticamente el mismo, con solo algunos cruces.
- *Alineamiento automático.* La información de alineamiento provista por GIZA cubre el 95 % de las palabras del idioma origen.

Los otros problemas observados en los experimentos 1 y 2 están todavía presentes: múltiples contornos con el mismo significado, la incidencia de la componente de grupo entonativo, y los problemas de consistencia entre idiomas.

Evaluación subjetiva

Con el objeto de obtener una medida subjetiva de la calidad de la entonación generada usando la información de la codificación de los contornos de entonación del hablante origen, se llevó a cabo una evaluación subjetiva en la dirección catalán→español.

La evaluación consistió en solicitar a los participantes calificar en una escala de uno a cinco el grado de naturalidad de 15 audios. Los audios corresponden a habla real del locutor, y audios resintetizados usando los contornos de entonación predichos usando tanto el algoritmo base como el propuesto. En total, cada participante escuchó cinco audios de cada condición: natural, base y propuesta.

En la Figura 5.10 se puede observar el *box-plot* de las evaluaciones correspondientes a las diferentes condiciones experimentales. El habla natural recibió el valor más alto de naturalidad, tal como era de esperar.

En la gráfica se observa que la utilización de la entonación del hablante origen permite obtener una mayor naturalidad de la entonación, ya que prácticamente en el 75 % de los casos es superior a la mediana del algoritmo base.

Un análisis usando el *Mann-Whitney-Wilcoxon test* permite establecer que con $p < 0{,}01\,\%$ tanto la distribución de los valores de naturalidad de los algoritmos base como del propuesto es diferente a la del habla natural. En consecuencia, aunque el algoritmo propuesto es superior al algoritmo base, todavía los oyentes pueden percibir una diferencia en naturalidad con respecto al habla del locutor original.

Según el test, la distribución de los valores de MOS del algoritmo propuesto es diferente a la correspondiente al algoritmo base con $p < 8\,\%$. En consecuencia, se puede observar una tendencia en lo referente a la superioridad del algoritmo propuesto sobre el algoritmo base, tal como lo reflejaban los resultados objetivos.

En la Figura 5.11 se puede observar el contorno predicho para diferentes condiciones: usando solamente información lingüística, usando la codificación del contorno de entrada, y usando ambas. La frase es: "Le puedo decir para las fechas que usted pide... temporada alta... veamos... en lo que sería pensión completa... el precio por persona y noche es de mil trescientos euros."

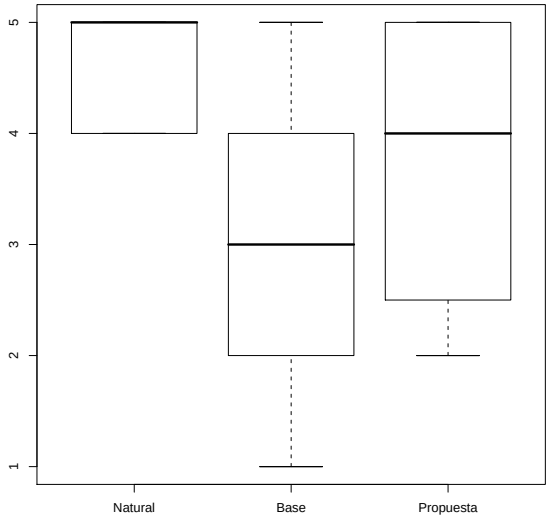


Figura 5.10: MOS de naturalidad obtenido para las diferentes condiciones experimentales usando los datos de evaluación.

5.4. Generación de la duración utilizando información de la fuente

Como se mencionó en la Sección 2.3, existe una serie de fuentes de variación de la duración de los segmentos. Estas fueron clasificadas por Klatt en factores segmentales, silábicos y suprasilábicos. Allí también se indicó que otro factor importante es la velocidad del habla [Kla76].

En los factores segmentales, silábicos y suprasilábicos interfieren fenómenos articulatorios, fonológicos, lingüísticos y paralingüísticos. Los tres primeros son intrínsecos del idioma y pueden ser modelados por los algoritmos descritos en la Sección 2.3, la cual describe diversos modelos de duración.

Por otra parte, los fenómenos paralingüísticos son difíciles de predecir. Son elementos no verbales de la comunicación usados para modificar el significado, expresar una emoción, indicar una intención, etc. En el caso de la duración los fenómenos paralingüísticos se ven en parte reflejados en el cambio de ritmo, y como consecuencia, la elisión de algunos fonemas en el caso de un aumento de la velocidad del habla.

En la siguiente Sección se hará una introducción sobre la influencia del ritmo en las diferentes unidades del habla (sílaba, palabra y oración), para luego analizar en la Sección 5.4.2 acerca de la posibilidad de transferir información del ritmo de un idioma al otro. Posteriormente, en la Sección 5.4.3 se estudiará un enfoque orientado a la sincronización de video del locutor en el idioma origen y el audio traducido sintetizado usando conversión texto a voz. En algunos aspectos esto implicará una transferencia indirecta del ritmo entre

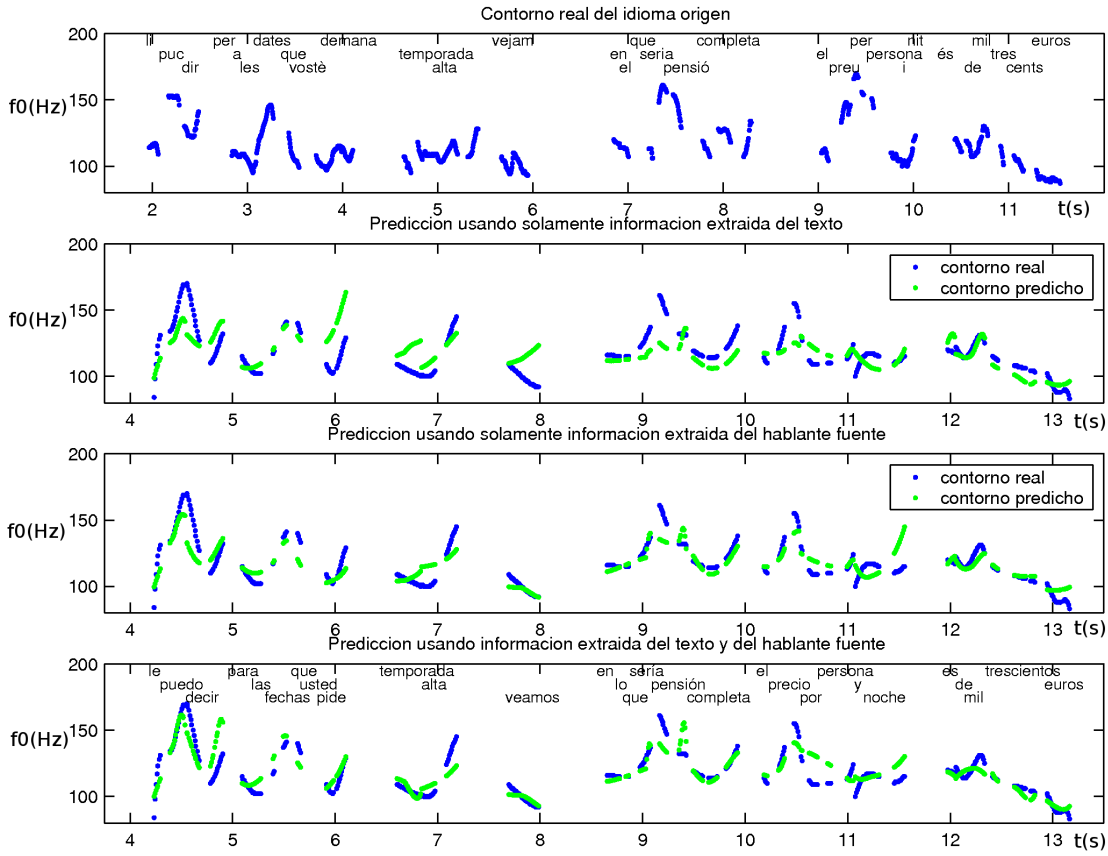


Figura 5.11: Ejemplo de contornos predichos usando tanto información lingüística como la codificación del contorno de entrada.

idiomas.

5.4.1. Influencia del ritmo en las unidades del habla

Los cambios en la velocidad del habla se ven reflejados en distintos niveles: sílaba, palabra y oración. Por ejemplo, cuando una persona cambia la velocidad del habla, las duraciones de las sílabas no acentuadas en palabras polisilábicas se reducen en mayor medida que en el caso de las sílabas acentuadas [Jan04].

Existe diversas teorías acerca del motivo de los cambios introducidos a nivel segmental. Una de las posturas supone que los hablantes preservan aquellas partes de información del habla que son informativas [Lin90]. Otra de las posturas afirma que los cambios a nivel de sílaba, palabra u oración son debidos a restricciones en la articulación.

Un ejemplo que apoyaría esta última afirmación es el trabajo de Cho [Cho01], que encontró que las vocales acentuadas muestran una resistencia articulatoria más grande a los cambios que las vocales no acentuadas. Esto significa que el hablante es forzado a aplicar más energía en aproximar objetivos específicos relacionados con las sílabas acentuadas que

para aquellas que no lo son. En la literatura se pueden encontrar otros estudios que relacionan la velocidad de locución y la coordinación articulatoria: [Ada93][Sha95][Oka99][Oka03], entre otros.

En la tesis de Siegles [Sie95], se menciona que en los muchos estudios acerca de la influencia de la velocidad de locución, existen múltiples evidencias que afirman que las vocales sufren mayores cambios que las consonantes, dependiendo esto del tipo de vocal, sus vocales adyacentes y el tipo de palabra. En este punto es importante remarcar que las personas tienen dificultad para identificar fonemas rápidos cuando se insertan en habla a velocidad normal, demostrando la importancia de la adaptación a la velocidad de locución.

En esta tesis se propone la transferencia del ritmo entre idiomas en el marco de la traducción voz a voz, con dos objetivos específicos: la imitación del ritmo del hablante origen para mejorar la transferencia del estilo, y la sincronización del audio y del video para permitir que cierta información visual (como por ejemplo, lenguaje corporal) complemente la información auditiva, y además evitar desfasajes entre los diferentes canales de comunicación.

En la literatura se puede encontrar que la producción de habla comprimida está íntimamente relacionada con los objetivos de esta sección: manipular el ritmo de la persona sin afectar su inteligibilidad y reducir en forma controlada la duración de una locución sin afectar el ritmo, a través de la manipulación de las pausas. Covell et al. [Cov98] proponen el algoritmo Mach1, que se basa en imitar las estrategias de compresión que las personas utilizan cuando hablan rápido [San94][Wit93]:

- Compresión máxima de pausas y silencios.
- Compresión mínima de las vocales acentuadas.
- Compresión intermedia de las vocales no acentuadas.
- Compresión de las consonantes basada en los niveles de acentuación de las vocales adyacentes.
- Comprimir en promedio más las consonantes en lugar de las vocales.

Otro trabajo relacionado con el de Covell es el de He et al. [He00][He01], en donde se realiza un estudio comparativo de dos algoritmos. El primero de ellos utiliza una compresión lineal de los segmentos de habla, y se suprimen o acortan las pausas que fueron detectadas como silencios en el audio. Con la supresión de pausas se consiguen compresiones del 10 % al 25 % [Gan88]. A pesar de su simplicidad, este algoritmo presenta una aceptabilidad similar al algoritmo propuesto por Covell. El segundo algoritmo, basado en Mach1, presenta una reducida ventaja con respecto al otro algoritmo de mayor simplicidad, y los autores concluyen que no es recomendable ese aumento de carga de procesamiento.

Tucker et al. [Tuc00] realizaron un estudio más exhaustivo de técnicas de compresión de audio. No solo se incluye la compresión del habla y la reducción de pausas, sino que se analiza la técnica de *escisión*. Esta última consiste en eliminar silencios, *fillers*, palabras y oraciones sin importancia de las grabaciones [Sti96][Aro97] [Hor03] [McK05]. Los resultados demostraron que las técnicas de escisión eran mejores que la variación de la velocidad

del habla. Sin embargo, se detectó que esta técnica afecta la comprensión, llevando a los oyentes a una pérdida del contexto, y la necesidad de reproducir porciones cercanas del audio para lograr entender el contenido.

En esta tesis se propone combinar el enfoque de compresión (o expansión) a través de la manipulación de las pausas y el procesamiento lineal de los segmentos teniendo en cuenta la estructura rítmica de los mismos usando bien la isocronía silábica o bien la acentual.

5.4.2. Transferencia del ritmo entre idiomas.

El objetivo de esta sección es analizar si es posible transferir información del ritmo de un idioma al otro. Por ejemplo, si en el idioma origen (con isocronía silábica) se observa una velocidad promedio de las sílabas entre dos pausas de un valor determinado, estudiar si es posible transferir este valor al idioma destino a través de un factor de escala que afecte a la isocronía del mismo. De esta manera, el oyente percibiría las variaciones de ritmo.

Una vez conocida la duración de la unidad isocrónica (silábica o acentual), sería posible estimar la duración de cada fonema constituyente. Para ello se haría uso del algoritmo descrito en la Sección 3.2.1, el cual, a través de distintos pesos, permite distribuir la duración de la unidad isocrónica en sus fonemas constituyentes.

Para estudiar esta posibilidad se utilizó el corpus bilingüe del proyecto TC-STAR que consiste en cuatro hablantes bilingües: español e inglés británico. La segmentación de los audios en fonemas es automática, usando RAMSES [Bon98] en el modo de alineamiento forzado.

La información sobre alineamiento entre los idiomas se obtuvo usando GIZA++ [Och03], utilizando un corpus más grande e incluyendo la información que se quería alinear dentro del mismo. De esta manera se obtuvieron los vínculos entre palabras de los idiomas.

La Figura 5.12 contiene el gráfico de dispersión del ritmo en sílabas o acentos por segundo para el español y para el inglés británico, para cada una de las palabras del corpus que poseen un alineamiento entre idiomas. De esta manera, se puede estudiar la linealidad de la relación entre el ritmo de los diferentes idiomas para palabra correspondientes.

El ritmo para cada palabra en sílabas por segundo se calculó como el cociente entre el número de sílabas y el tiempo transcurrido entre dos pausas. En consecuencia, el ritmo estudiado es una medida global entre dos pausas, y a cada palabra entre pausas se le asignará el mismo valor de ritmo.

En el caso del ritmo en acentos por segundo para cada palabra se calculó de manera similar al ritmo de sílabas por segundo. Se realizó el cociente entre el número de acentos y el tiempo transcurrido entre dos pausas. Aquí también a cada palabra entre pausas se le asignará el mismo valor de ritmo de acentos por segundo.

Los cuatro gráficos de la izquierda representan la dispersión para cada uno de los locutores en el estudio de la relación lineal entre el ritmo en acentos por segundo para el español y el inglés británico. Los gráficos de la derecha representan la dispersión en el caso de utilizar el ritmo en sílabas por segundo para el español y el inglés británico.

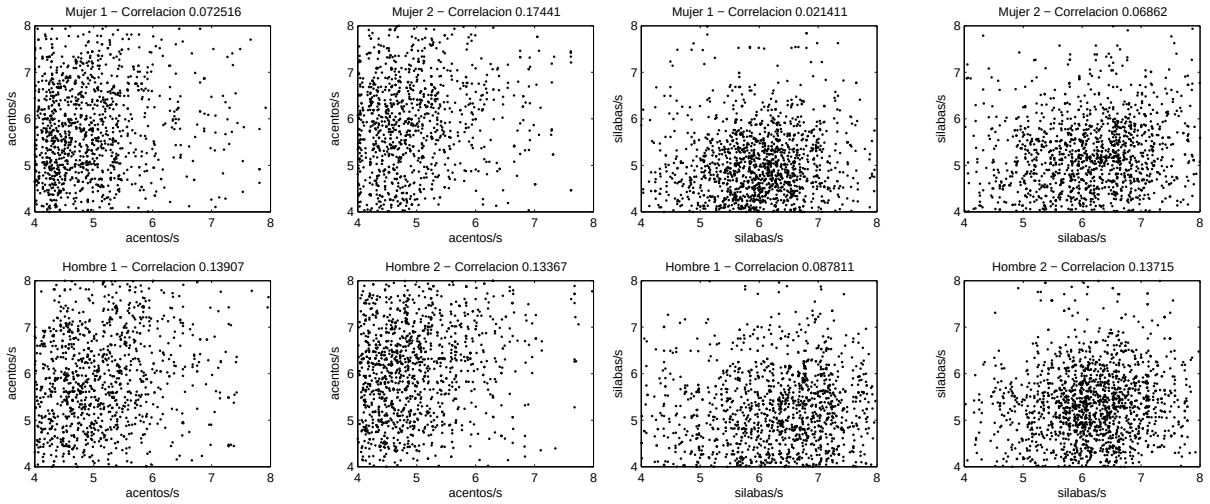


Figura 5.12: Dispersión del ritmo en sílabas y acentos por segundo para el español e inglés británico

Como se puede observar en la parte superior de cada gráfico, las medidas de ritmo para el español y el inglés no parecen estar correladas. El valor más grande de correlación asumiendo una relación lineal es 0,17, resultando muy pequeño.

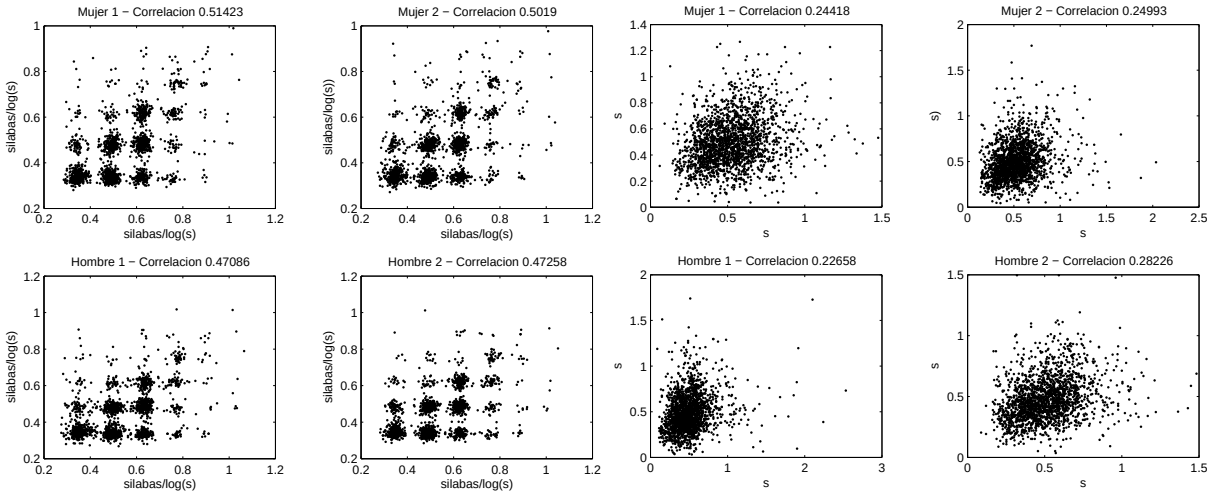


Figura 5.13: Correlación entre la velocidad de locución de los idiomas midiendo el ritmo a nivel de palabra usando el logaritmo de la duración de la misma

Figura 5.14: Correlación entre la duración de las palabras de los idiomas

La Figura 5.13 muestra mediciones de ritmo a nivel de palabra para ver la correlación entre la velocidad de locución de los idiomas, comparando el ritmo de palabras alineadas. En el gráfico de la izquierda se puede observar la correlación entre la velocidad de locución de los idiomas midiendo el ritmo a nivel de palabra haciendo el cociente entre el número de sílabas y el logaritmo de la duración de la misma. En el gráfico de la derecha se observa la

correlación entre la duración de las palabras de los idiomas usando la duración en segundos de las mismas.

A pesar de que la correlación ha aumentado, es todavía baja para resultar útil a los propósitos de la transferencia de ritmo entre idiomas. El aumento de la correlación, que llega a alcanzar valores de 0,51, se debe principalmente a la influencia del número de sílabas en las palabras en los dos idiomas. A causa de que existe un gran número de veces en las que el número de sílabas entre los dos idiomas coincide o bien solo difiere en uno, el coeficiente de correlación resulta cercano a 0,5.

En la Figura 5.14 podemos ver que la correlación entre las duraciones de las palabras en los idiomas es muy bajo, lo que indica que la correlación alcanzada en la Figura 5.13 se debe principalmente al efecto de la diferencia en el número de sílabas entre las palabras de los dos idiomas.

Como conclusión, de acuerdo a los datos disponibles en el proyecto TC-STAR, no es posible establecer una relación lineal entre los idiomas para ninguna de las medidas del ritmo. Esto ocurre a pesar del requerimiento realizado a los locutores sobre respetar la consistencia de estilos en la locución en los dos idiomas, ya mencionado en la Sección 5.3.1.

En consecuencia, no será posible utilizar esta información para transferir la velocidad del habla de un idioma a otro usando medidas del ritmo globales (entre pausas) ni locales (a nivel de palabra).

5.4.3. Sincronización de los audios de dos idiomas.

Una aplicación de la traducción oral es la traducción de documentos multimedia, conteniendo audio y video. Este es el caso de la traducción de programas de televisión como canales parlamentarios, de noticias, etc. Para obtener una salida adecuada es necesario usar técnicas de sincronización, de lo contrario podría escucharse información que no se corresponde con lo que se está mostrando en el video. De esta manera, en forma indirecta, se produce una transferencia del ritmo del hablante origen en el idioma destino.

La sincronización entre los audios de los dos idiomas se realiza a nivel de palabra (tiempos de referencia a sincronizar), utilizando información de alineamiento provista por el sistema de traducción automática. En consecuencia, el objetivo será lograr la mejor sincronización posible en el tiempo de inicio de la pronunciación de una palabra mediante el conversor texto a voz en el idioma traducido (luego de haberse realizado la traducción mediante SMT), y el instante en que en el video se pronuncia la misma palabra en el idioma origen.

Con el fin de lograr una adecuada sincronización, se decide preservar la monotonía creciente de los tiempos usados como referencia para sincronizar. En consecuencia, todos aquellos alineamientos que provoquen cruces y alteren la monotonía creciente no serán utilizados. Por ejemplo, en la Tabla 5.1 se observa que el tiempo marcado en negrita no preserva la monotonía creciente en el idioma origen, y por lo tanto, no será utilizado en la sincronización. Por otra parte, a pesar de que el tiempo marcado en cursiva preserva la monotonía creciente en el idioma origen, no será utilizado en la sincronización debido a que las palabras entre idioma origen y destino no corresponden.

Palabra fuente	**Audio fuente**	**Palabra destino**	**Audio destino**
PALABRA 1	4.38	PALABRA 1	2.92
PALABRA 2	5.28	PALABRA 2	3.55
PALABRA 3	**2.09**	**PALABRA 4**	**4.27**
PALABRA 4	*6.37*	*PALABRA 3*	*4.86*

Tabla 5.1: Selección de los tiempos de referencia para la sincronización con el objeto de mantener una monotonía creciente.

5.4.3.1. Compresión de pausas

La sincronización se ha realizado utilizando dos algoritmos. El primero de ellos solamente realiza compresión o expansión de pausas tal como han propuesto muchos autores para la compresión de audio, ya tratados en la Sección 5.4.1. Este algoritmo se aplica en forma cronológica, es decir, en orden creciente del tiempo. Por lo tanto, también resulta útil en aplicaciones en tiempo real con pequeños tiempos de retardo.

En la Tabla 5.2 se puede observar la presencia de una pausa, y se puede considerar que la misma posee una correspondencia entre idiomas, ya que no se presentan alineamientos entre palabras anteriores y posteriores a la pausa en ambos idiomas. Aquellas pausas que reúnen estos requisitos serán las usadas para realizar la sincronización.

Palabra fuente	**Audio fuente**	**Palabra destino**	**Audio destino**
PALABRA 1	4.38	PALABRA 1	2.92
PALABRA 2	5.28	PALABRA 2	3.55
PALABRA 3	2.09	PALABRA 4	4.27
PALABRA 4	6.37	PALABRA 3	4.86
PAUSA	8.42	PAUSA	6.22
PALABRA 5	8.92	PALABRA 5	6.53

Tabla 5.2: Selección de los tiempos de referencia para la sincronización usando pausas.

Los resultados de aplicar esta técnica se presentan en la Figura 5.15. Aquí se muestra un histograma del error en el tiempo de sincronización, es decir, la diferencia entre el tiempo inicial del audio sintetizado usando TTS y el tiempo de referencia obtenido en el audio origen, para los cuatro locutores del corpus.

La sincronización de audio utilizando únicamente eliminación de pausas presenta problemas debido a que los audios resultan en media retrasados 400 milisegundos. Esto es consecuencia de la falta de tiempo de pausas suficiente para lograr la perfecta sincronización.

5.4.3.2. Compresión de ritmo

El segundo algoritmo, además de alterar la duración de las pausas, cambia el ritmo entre tiempos de sincronización (pausas con correspondencia entre idiomas) dentro del límite del 10 % de la duración original de las mismas, para no provocar cambios bruscos de velocidad que no serían naturales. Por ejemplo, dada una porción de audio entre pausas

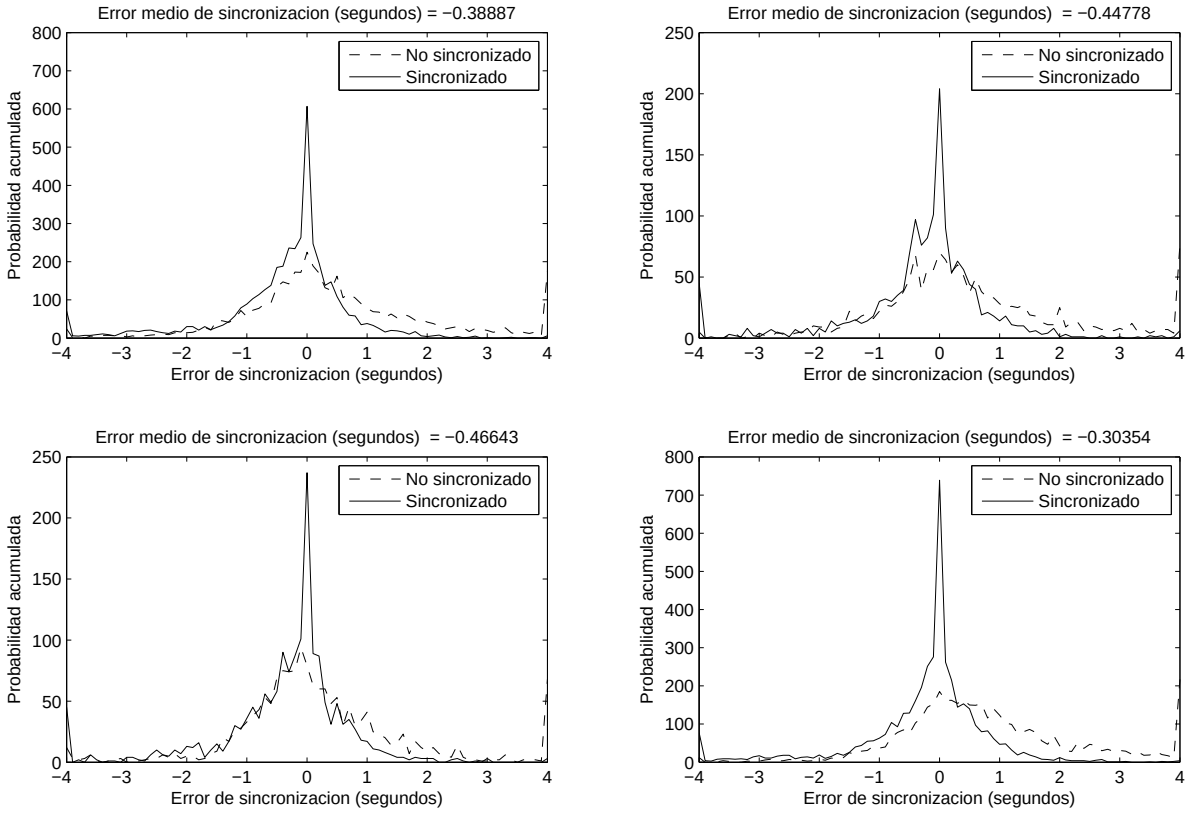


Figura 5.15: Precisión de la sincronización utilizando algoritmos de compresión/expansión de pausas.

cuya duración sea de 1,5 segundos, su duración se podría expandir o contraer como máximo en ±150 milisegundos distribuidos de manera uniforme entre las sílabas constituyentes.

En la Figura 5.16 se muestran los resultados de sincronización para cada uno de los cuatro locutores utilizando tanto compresión/expansión de pausas como alteración reducida del ritmo. El uso de esta técnica reduce el error de sincronización cerca de 250 milisegundos con respecto a la técnica de compresión de pausas. El error de sincronización queda reducido a menos de 120 milisegundos.

Resultados informales con documentos audiovisuales han demostrado que el uso de estas técnicas son del agrado de la audiencia en diversas demostraciones, debido a que se observaba el estilo del hablante origen y también una buena sincronía con los otros canales de comunicación, como por ejemplo los gestos, lo cual no es común en muchas traducciones.

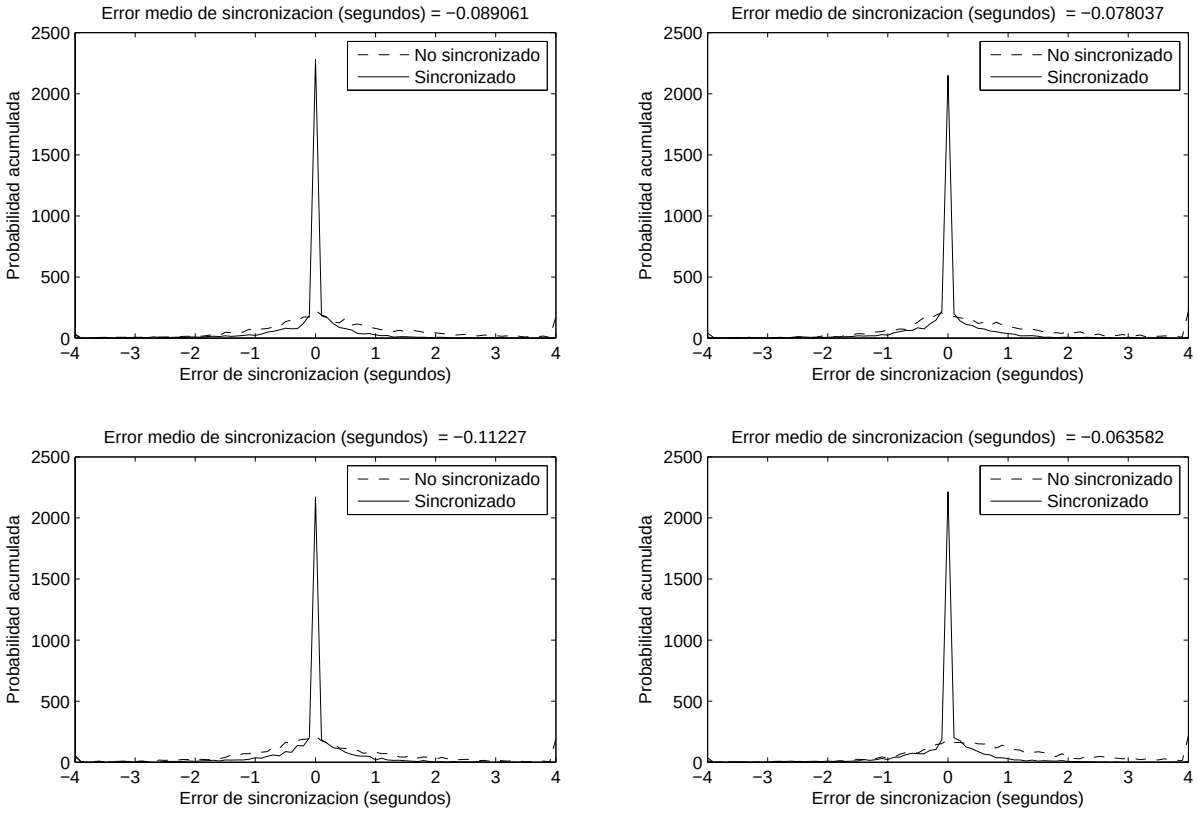


Figura 5.16: Precisión de la sincronización utilizando algoritmos de compresión/expansión de pausas y segmentos de voz.

5.5. Generación de pausas usando información de la fuente

En el marco de la traducción voz a voz resultaría de interés el uso de las pausas del hablante en el idioma origen para trasladar parte de su estilo y ofrecer más expresividad en el idioma destino. Esta enfoque contribuye a convertir la traducción voz a voz en completa, abarcando desde el contenido, su forma de expresarlo a través de la prosodia, hasta llegar a incluir la identidad de la voz de salida, usando técnicas de conversión de voz. Además, tales pausas permitirán evitar malentendidos o ambigüedades introducidas por otras técnicas de predicción, tal como se explica en la Sección 2.4.

En la literatura se encuentran algunos ejemplos sobre el uso de las pausas en el contexto de la traducción voz a voz. En este marco, Verbmobil fue un proyecto que contribuyó a la utilización de la prosodia en todo el proceso de la traducción voz a voz, tanto en el procesamiento del dialogo y su traducción, como en la generación de la síntesis de voz. Entre las diversas características prosódicas analizadas se incluyó la utilización de la información relativa a la presencia de pausas y su duración al comienzo y al final de cada palabra [Nie97]. Su utilización contribuyó a resolver el problema de la falta de puntuación,

detectando importante información semántica de una oración y mejorando la traducción [Blo97].

Otros ejemplos del uso de las pausas se encuentran en la traducción automática estadística, donde numerosos autores las han utilizado (aquellas cuya duración es mayor a 300ms) para obtener segmentos más pequeños a traducir (aproximadamente 10s) [Gal07]. Además, las pausas cumplen un importante papel en el agregado de la puntuación [Liu04a][Hua02][Kim01] para mejorar la calidad de la traducción [Füg07]. La importancia de la puntuación, junto con otra información sobre tipografía (mayúsculas/minúsculas), segmentación en oraciones y normalización de palabras, permitieron al sistema de LIMSI/CNRS superar el mejor sistema en la evaluación 2007 de TC-STAR por casi 2 puntos BLEU [Déc07].

En esta tesis se propone realizar la transferencia de las pausas utilizando la información provista por cada uno de los componentes del sistema de traducción voz a voz: ASR, SMT y TTS.

En el caso del ASR se utilizará la información de las pausas presentes en el idioma fuente, que fueron detectadas usando diferentes modelos acústicos del silencio. Teniendo esta información, es posible intentar la transferencia de estas pausas al idioma destino utilizando información sobre el alineamiento presente en el componente de traducción automática.

En la próxima sección se explicará el algoritmo propuesto para la transferencia de las pausas del idioma origen al destino, y su combinación con los algoritmos explicados en la Sección 3.3 sobre generación de junturas terminales.

5.5.1. Transferencia de pausas usando tuplas.

En esta tesis se propone hacer uso de las tuplas para la transferencia de pausas entre idiomas. En traducción las tuplas son unidades bilingües, que se pueden definir formalmente como el conjunto de las frases más cortas que proporcionan una segmentación monótona de los datos bilingües. El conjunto de reglas a seguir para extraer las tuplas de cualquier alineamiento palabra a palabra, que puede incluir cambios de orden, son las siguientes [Cre04]:

- Se debe producir una segmentación monótona de cada par de oraciones bilingües.
- Ninguna palabra dentro de una tupla se alinea con palabras fuera de la tupla.
- No es posible extraer ninguna nueva tupla dentro de una más grande sin violar las dos reglas anteriores.

Un ejemplo de tuplas se puede observar en la oración “La casa blanca”, donde las tuplas serían [LA]-[THE] y [CASA BLANCA]-[WHITE HOUSE]. Como consecuencia de estas reglas, existe solamente una segmentación posible para cada par de oraciones bilingües. Nótese que debido al procedimiento de alineado, pueden aparecer algunas tuplas

que consisten en un fragmento monolingüe sin alineamiento con ninguna palabra del otro idioma, convirtiéndose en tuplas sin información útil para nuestros propósitos.

El análisis de datos bilingües reveló que la mayoría de las pausas en la frontera de una tupla en el idioma origen se pueden transferir directamente a la frontera de la tupla correspondiente en el idioma destino. Un ejemplo se observa en el siguiente párrafo extraído de la base de datos utilizada:

> [Before I do, however], PAUSA [I want] [to outline] [the] [two] [key principles] PAUSA [which underlined] [ireland 's presidency] PAUSA [and] [indeed] [underpin] [ireland 's general approach] PAUSA [to] [european affairs]. PAUSA
>
> [No obstante, PAUSA antes] [quisiera] [esbozar] [los] [dos] [principios básicos] PAUSA [subyacentes] [a la presidencia irlandesa] PAUSA [y] [que de hecho] [sustentan] [el enfoque general de irlanda] PAUSA [con respecto a] [los asuntos europeos]. PAUSA

En este ejemplo se puede observar que la primera pausa se encuentra dentro de una tupla, y por lo tanto no podrá ser considerada para la transferencia. El resto de las pausas si se encuentran en la frontera de una tupla, y pueden ser transferidas.

Una de las consecuencias importantes de la utilización de tuplas es la reducción de los efectos de reordenamiento en la transferencia de pausas. Por ejemplo, en la oración *The White House PAUSA*, la pausa después de *House* sería transferida a *La Casa PAUSA Blanca* si se asociara la pausa a la palabra anterior. Este resultado es incorrecto, ya que la pausa está dentro del sintagma nominal en el idioma destino, mientras que esta en la frontera del mismo en el idioma origen. Sin embargo, las tuplas proporcionan una transferencia de pausas correcta debido a que *White House* y *Casa Blanca* es una tupla, y por lo tanto la pausa se transfiere a la frontera de la tupla como se deseaba: *La Casa Blanca PAUSA*.

Sin embargo, una importante limitación aparece cuando una pausa cae dentro de un grupo de palabras con alineamientos "muchos a muchos". En este caso no es posible encontrar la posición de la pausa en forma precisa en el idioma destino, ya que muchas palabras en el idioma destino tienen un alineamiento con la palabra previa a la pausa en el idioma origen. El mismo efecto ocurre cuando hay un alineamiento faltante.

Los algoritmos de predicción de junturas terminales de la Sección 2.4 se utilizarán para predecir estas pausas faltantes, teniendo en cuenta que muchas de ellas ya están predichas debido a que fueron transferidas usando el algoritmo por tuplas. De esta manera, las pausas resultantes serán diferentes que aquellas producidas por el sistema que no hace uso de la transferencia de la información del hablante origen.

Reutilizando el ejemplo anterior, la pausa marcada en negrita en el idioma español no puede ser transferida a la pausa marcada en negrita en el idioma inglés. Dicha pausa en el idioma inglés deberá ser predicha usando los algoritmos de predicción de junturas terminales de la Sección 2.4.

> [Before I do, however], **PAUSA** [I want] [to outline] [the] [two] [key princi-

ples] PAUSA [which underlined] [ireland 's presidency] PAUSA [and] [indeed] [underpin] [ireland 's general approach] PAUSA [to] [european affairs]. PAUSA

[No obstante, **PAUSA** antes] [quisiera] [esbozar] [los] [dos] [principios básicos] PAUSA [subyacentes] [a la presidencia irlandesa] PAUSA [y] [que de hecho] [sustentan] [el enfoque general de irlanda] PAUSA [con respecto a] [los asuntos europeos]. PAUSA

En la siguiente sección se muestran los resultados experimentales que indican las ventajas de este enfoque a través de estudios con medidas objetivas y subjetivas.

5.5.2. Condiciones experimentales

En estos experimentos se ha usado los datos generados para el proyecto TC-STAR que corresponden a cuatro hablantes bilingües (español e inglés británico). Las pausas se detectaron automáticamente usando el reconocedor automático del habla RAMSES [Bon98]. La información de alineamiento fue generada automáticamente usando GIZA++ [Och03].

Los datos se han dividido en diez partes para hacer experimentos de *10-fold cross validation*. Los resultados se presentan usando varias métricas para estudiar el rendimiento de los sistemas: precisión, recall y F-measure (tal y como se utilizó en la Sección 4.4.2).

Los experimentos se realizaron usando un sistema base que no usa información del hablante origen (transductor de estados finitos), y otro sistema que implementa el enfoque propuesto (tuplas+transductor de estados finitos).

A pesar de que los hablantes fueron instruidos para ser consistentes prosódicamente en ambos idiomas, la relación de pausas entre idiomas no es una a una, sino que hay pausas que aparecen en un idioma y no en el otro. Debido a esto, se ha realizado un segundo análisis manual para estudiar la calidad de las pausas predichas. Es sabido que las medidas objetivas mencionadas anteriormente tienen tendencia a ser pesimistas debido a que solamente se posee una referencia con la cual comparar la precisión de la predicción.

5.5.3. Resultados experimentales

Los resultados del sistema base se muestran en las filas como FST de la Tabla 5.3, mientras que en las líneas Tuplas+FST se detallan los resultados de la propuesta. También se incluyeron los resultados obtenidos mediante la utilización de tuplas sin las pausas introducidas por FST, y se encuentran en las líneas del algoritmo Tuplas.

El algoritmo propuesto tiene una menor precisión con respecto al sistema base debido al mayor número de pausas predichas. Existen un 20 % de pausas adicionales con respecto a las predichas por el algoritmo base.

El recall muestra que las pausas predichas son mejores que las generadas con el sistema base, lo que es un indicador de que muchas pausas están mejor predichas utilizando el algoritmo propuesto que a través del uso de FST en forma aislada.

El algoritmo Tuplas tiene una mejor precisión que Tuplas+FST, ya que hay un mayor número de pausas correctas dentro de las pausas predichas. Sin embargo, el recall de Tuplas es menor que Tuplas+FST debido a que faltan muchas pausas debido a que no fue posible transferirlas porque se encontraban dentro de una tupla

Locutor	Algoritmo	Precision	Recall	F-Measure
Mujer 1	FST	58.17	65.29	61.41
Mujer 1	Tuplas	62.09	42.78	50.59
Mujer 1	Tuplas+FST	54.82	70.58	61.62
Mujer 2	FST	60.02	67.49	63.49
Mujer 2	Tuplas	58.64	48.97	53.34
Mujer 2	Tuplas+FST	54.44	75.17	63.11
Hombre 1	FST	61.10	66.12	63.37
Hombre 1	Tuplas	58.65	45.73	51.16
Hombre 1	Tuplas+FST	54.83	73.15	62.57
Hombre 2	FST	58.88	64.27	61.32
Hombre 2	Tuplas	58.41	44.34	50.36
Hombre 2	Tuplas+FST	53.25	72.22	61.23

Tabla 5.3: Resultados experimentales para los diferentes enfoques usando una comparación objetiva con una referencia.

En algunos casos las pausas en el idioma origen no están ubicadas en la posición correspondiente en el idioma destino. El hablante bilingüe ha tomado una decisión diferente, introduciendo una inconsistencia en los datos paralelos. Un análisis manual de las pausas transferidas usando tuplas que no se encuentran en el hablante destino (falsos positivos) revela que un 83 % de ellas se encuentran en la ubicación correcta, debido a que el locutor utilizó una pausa en el idioma origen y decidió no hacerla en el idioma destino.

Si se consideran que todas las pausas transferidas son correctas, lo cual es erróneo en el 17 % de los casos, los valores de las medidas objetivas se ven alterados como se muestra en la Tabla 5.4. Estos resultados son una cota superior al máximo rendimiento alcanzable con la utilización de la transferencia de pausas usando tuplas, utilizando el método FST para la predicción de pausas.

Locutor	Algoritmo	Precision	Recall	F-Measure
Mujer 1	Tuplas+FST	75.00	76.59	75.70
Mujer 2	Tuplas+FST	79.50	81.44	80.39
Hombre 1	Tuplas+FST	78.99	79.48	79.17
Hombre 2	Tuplas+FST	76.67	78.72	77.61

Tabla 5.4: Resultados experimentales para los diferentes enfoques usando una comparación objetiva con una referencia, considerando que todas las pausas transferidas son correctas.

A continuación se muestra un ejemplo de un párrafo paralelo. El idioma origen es el inglés y el idioma destino es el español. En los párrafos del idioma origen se han detectado pausas (simbolizadas como <S>) del hablante usando ASR, mientras que las pausas del

párrafo destino fueron predichas:

Grabación en inglés: "... We have clarified the division of powers <S> between the union <S> and the member states. <S> It is now clear <S> how decisions are taken, <S> and who is entitled to take them. ..."

Traducción al español con pausas predichas mediante el método tuplas+FST: "... Hemos aclarado la división de poderes <S> entre la unión <S> y los estados miembros. <S> Ahora queda claro <S> cómo se adoptan las decisiones <S> y quién está autorizado a adoptarlas. ..."

Grabación paralela en español: "... Hemos aclarado la división de poderes entre la unión y los estados miembros. <S> Ahora queda claro cómo se adoptan las decisiones <S> y quién está autorizado <S> a adoptarlas. ..."

Es de destacar la precisión con la que se transfieren las pausas del hablante origen en el ejemplo, lo cual contribuye a una mejora en la expresividad del conversor texto a voz. Además, existe una adaptación al estilo de locución del hablante origen no presente en muchas de las traducciones realizadas por intérpretes humanos.

El análisis manual de los pausas transferidas revela que en muchos casos el error cometido surge de problemas de alineamiento. En algunas ocasiones las tuplas en un idioma abarcan palabras que no se corresponden con las contenidas en la tupla del otro idioma. Por ejemplo, en el siguiente fragmento de oración se encuentran las tuplas delimitadas por corchetes:

"... [Mister] [President,]<S> [Madam] [Vice President,] <S> [commissioners] ..."

"... [Señor] [Presidente,] <S> [Señora] [Vicepresidenta y señores] <S> [comisarios] ..."

Debido a los enlaces erróneos proporcionados por Giza++, la tercer tupla del idioma destino contiene palabras adicionales con respecto a la tupla correspondiente del idioma origen. Como consecuencia de esto, la pausa es transferida de manera errónea luego de la palabra **señores**, cuando dicha palabra debería estar dentro de la tupla [**y señores comisarios**].

5.6. Conclusiones

En el presente capítulo se han tratado diversos temas relacionados con la prosodia en el marco de la traducción voz a voz, tales como la transferencia de la entonación, el ritmo y las pausas.

En la Sección 5.3 se estudió la transferencia de la entonación de un idioma a otro. Para ello se consideró la posibilidad de utilizar esquemas de anotación existentes, tales como ToBI o INTSINT. De esta manera, una vez obtenida la anotación de ambos idiomas, sería posible aplicar técnicas de aprendizaje automático para encontrar relaciones entre las anotaciones. Sin embargo, la conclusión fue que en este tipo de esquemas de anotación de eventos tonales se realizan ciertas suposiciones, tales como una discretización taxativa de los contornos, que pueden forzar el ajuste del fenómeno al esquema de anotación, y

no viceversa, que es lo deseado. Esto pueden llevar a una anotación deficiente de los eventos tonales, y la utilización de esta información errónea solo conduciría a resultados pobres en la transferencia de la entonación. En consecuencia, se decidió la utilización de un enfoque de agrupamiento automático que permita encontrar un cierto número de tipos de movimientos tonales relacionados en los dos idiomas sin utilizar ninguna suposición acerca de su número. De esta manera, es posible utilizar esta codificación (obtenida luego del agrupamiento automático) de los contornos tonales del idioma origen como característica adicional en el modelado de la entonación del idioma destino. Para este objetivo se hizo uso de la segmentación en palabras del audio del idioma origen y destino dada por el ASR, contornos de frecuencia fundamental de los audios calculados por un algoritmo de extracción e información de alineamiento proporcionada por la traducción automática (GIZA++). Se han propuesto dos algoritmos de agrupamiento similares, presentando una diferencia importante en el aspecto de la clasificación de los movimientos como relacionados o no. El objetivo era encontrar un algoritmo de agrupamiento que permitiera encontrar movimientos tonales relacionados, sin perder capacidad de generalización. Los resultados experimentales demostraron la mejora introducida en el modelado de la entonación debido al enfoque propuesto, en comparación con un sistema base que no utiliza la información de la codificación del contorno del idioma origen. La mejora es importante en idiomas cercanos, tales como español y catalán. En el caso del español y el inglés, los resultados son sólo ligeramente mejores, debido en parte al origen diferente de los idiomas: latino y germánico respectivamente.

La transferencia del ritmo ha sido otro de los temas tratados en este capítulo (Sección 5.4). Se ha propuesto un método que combina la transferencia del ritmo y la sincronización entre audios. Este último aspecto fue considerado debido al uso de la tecnología de traducción voz a voz en conjunción con video. Coordinar los aspectos gestuales con la voz traducida es importante a causa de los múltiples canales involucrados en la comunicación humana. El algoritmo de sincronización incluye en su funcionamiento diversas características propias de algoritmos de compresión de audio, tales como compresión de pausas y silencios, y compresión por modificación del ritmo, utilizados en el algoritmo Mach1 [Cov98]. Mediante experimentos utilizando las características mencionadas se obtienen errores de sincronización muy bajos, cercanos a los 150 milisegundos, que lo convierte en apto para su uso en sincronización de audio/video.

Finalmente, en la Sección 5.5 se ha explicado una técnica de transferencia de pausas en el marco de la traducción voz a voz, mediante la utilización de información sobre alineamiento. El estudio de los datos de entrenamiento utilizando dos tipos diferentes de unidades de traducción, palabras y tuplas, arrojó como resultado la ventaja del uso de esta última para la transferencia de pausas. La tupla permite agrupar es su interior palabras que presentan un ordenamiento diferente entre idiomas. En consecuencia, es posible transferir las pausas de un idioma a otro usando la existencia de pausas en la frontera de las tuplas. Una limitación importante de este enfoque es la imposibilidad para trasladar una pausa de una tupla de un idioma a otro, si esta se encuentra dentro de la misma. El algoritmo compensa esta deficiencia realizando una predicción de pausas utilizando algoritmos convencionales (tales como los explicados en la Sección 2.4), teniendo en cuenta las pausas ya predichas mediante la transferencia de pausas entre idiomas.

Como conclusión podemos ver que la información del hablante fuente es útil para generar la prosodia del hablante destino en el proceso de traducción voz a voz. Sin embargo, se observó que en algunos casos solo es posible su utilización cuando los idiomas poseen un origen similar, tal como ocurre con el español y el catalán, siendo ambos lenguas latinas. Además, hay que tener en cuenta que los experimentos fueron realizados utilizando el mismo hablante tanto para el idioma origen como para el destino. El desarrollo de algoritmos independientes del hablante forma parte de las líneas futuras surgidas de la presente tesis.

Capítulo 6

Conclusiones y direcciones futuras

Uno de los objetivos generales de esta tesis fue investigar en el área del modelado de la prosodia en los sistemas de conversión texto a voz. Para ello se hizo un estudio detallado de la bibliografía involucrada, individualizando las distintas corrientes de opinión y enfoques, y detectando ciertos aspectos que podrían ser mejorados. Las contribuciones se enfocaron en el modelado de la entonación, la duración y las junturas terminales.

El segundo de los objetivos de la tesis fue el estudio de la transferencia de la prosodia en el marco de la traducción voz a voz, con el objeto de enriquecer la expresividad de la conversión texto a voz luego de la traducción automática. En este sentido se analizó una extensa bibliografía relacionada con el área, y también de otros temas que resultaron estar involucrados indirectamente. Los aspectos prosódicos incluídos en esta parte de la tesis fueron la entonación, el ritmo y las pausas. En esos aspectos es donde se enfocaron las diferentes contribuciones.

Generación de la prosodia

En el caso del modelado de la entonación se realizó una contribución para la eliminación de ciertas suposiciones inherentes de la mayoría de las metodologías de entrenamiento de la literatura, tales como requisitos de continuidad del contorno de frecuencia fundamental o la consistencia de la parametrización de los datos de entrenamiento.

El algoritmo propuesto (JEMA) combina los pasos de extracción de parámetros y generación del modelo en un bucle cerrado de mejora continua. En cada iteración se refinan tanto los parámetros como el modelo, obteniendo soluciones que aproximan de manera mas detallada el comportamiento de la entonación de los datos de entrenamiento.

Los resultados experimentales apoyan el enfoque propuesto, ya que todos los modelos de entonación estudiados (Bezier, Bezier superposicional y Fujisaki) presentan mejoras con respecto a un entrenamiento basado en los algoritmos de la literatura sobre el tema, donde la extracción de los parámetros y la construcción del modelo son pasos separados (SEMA). Los modelos que usan el enfoque propuesto alcanzaron un MOS de naturalidad en el rango de 3,2 a 3,5, y de calidad en el rango de 3,6 a 3,8. Mientras tanto, los modelos

entrenados con el enfoque SEMA recibieron un MOS de naturalidad en el rango de 2,3 a 3,1, y de calidad en el rango de 3,1 a 3,4.

JEMA fue estudiado también en muchos otros aspectos, los cuales han quedado fuera de la tesis. El enfoque propuesto fue aplicado a diversos idiomas, tales como español, catalán, inglés, esloveno, francés y chino mandarín. En todos ellos se obtuvieron mejoras con respecto a otros algoritmos encontrados en la literatura.

Con el objeto de corroborar las ventajas de JEMA sobre SEMA, se decidió realizar experimentos utilizando contornos y características lingüísticas artificiales. De esta manera se obtiene un conjunto de datos controlado que posee toda la información necesaria para el correcto modelado. La aplicación de JEMA a este conjunto de datos reveló su superioridad frente a SEMA para distintos niveles de ruido y ausencia de información (debido a segmentos sordos).

En lo relacionado con la generación de la duración se estudió el uso de la isocronía del idioma como característica principal para la predicción de la duración segmental. A través de un análisis de los datos de entrenamiento se demostró la dependencia entre la duración de la sílaba y el número de segmentos constituyentes. Como consecuencia de estas observaciones se propusieron un conjunto de modelos que permiten la predicción de la duración segmental en base a la duración suprasegmental. Mediante el uso del enfoque JEMA en el modelado de la duración se predicen conjuntamente la duración segmental y la suprasegmental, teniendo en cuenta las relaciones entre ellas.

Los experimentos indicaron la ventaja del enfoque propuesto sobre otros algoritmos de predicción de la duración segmental que no hacen uso de la información suprasegmental o del enfoque JEMA. Tener en cuenta la estructura rítmica del idioma proporcionada por las sílabas es importante desde el punto de vista perceptual, ya que los humanos somos capaces de detectar diferencias allí. Esto último ocurre, por ejemplo, cuando escuchamos hablar a un extranjero que no domina nuestro idioma y se perciben diferencias en la cadencia.

La evaluación objetiva de naturalidad y calidad reveló pequeñas diferencias entre los algoritmos estudiados que no resultaron estadísticamente significativas para $p = 0{,}05$ usando el *Mann-Whitney-Wilcoxon test*. Solamente uno de los algoritmos propuestos es estadísticamente mejor que un algoritmo base incluído en los experimentos, que no usa información de contexto para la predicción de la duración segmental.

Finalmente, en lo vinculado con el modelado de las junturas terminales, se hizo un estudio comparativo de algoritmos propuestos en la literatura y uno que utiliza transductores de estados finitos. Además, se evaluó la ventaja de usar el grupo acentual como unidad de análisis, en lugar de la palabra. El uso de esta unidad reduce el espacio de búsqueda y las posibilidades de colocar una juntura terminal dentro de un grupo acentual, hecho que fue observado como muy poco probable dentro de los datos de entrenamiento. Esto último se debe principalmente a que no se trata de habla espontánea, la cual presenta disfluencias que podrían originar la aparición de junturas dentro de un grupo acentual.

Los resultados revelaron que ninguno de los algoritmos estudiados tiene una precisión mejor que los otros, y su valor es cercano al 87 % para el hablante femenino y 78 % para el hablante masculino. Esto significa un alto porcentaje de junturas terminales predichas

correctamente sobre el total de junturas predichas para el hablante femenino. El resultado es importante, debido a que es necesario ubicar junturas terminales en los lugares correctos con poco errores.

En los resultados no se observaron grandes beneficios en la utilización del grupo acentual para la predicción de junturas terminales en lugar de la palabra. La única ventaja resulta en una disminución de la cantidad de decisiones que se deben tomar, reduciendo la carga computacional. Es de destacar que la ganancia es mínima debido a que la carga computacional es ínfima.

Transferencia de la prosodia

En esta tesis se abordó el tema de la transferencia de prosodia en el marco de la traducción voz a voz. El objetivo es utilizar la prosodia del hablante del idioma origen para mejorar la expresividad de la conversión texto a voz del texto traducido obtenido mediante traducción automática.

La Sección 5.3 consideró distintas alternativas para el uso de la entonación del hablante origen para enriquecer a aquella generada por el modelo de entonación. El enfoque considera la codificación del contorno de frecuencia fundamental del hablante origen, para luego ser utilizado (mediante la información de alineamiento) como característica adicional del modelo de entonación.

Luego del estudio de la factibilidad de distintos esquemas de codificación de la literatura, tales como ToBI e INTSINT, se decidió utilizar un enfoque de agrupamiento automático que presenta ciertas ventajas de adaptabilidad al dominio y ausencia de un conjunto reducido de clases que puede limitar la capacidad de modelado del sistema.

Aquellos idiomas mas cercanos debido a su origen latino, tales como español y catalán, presentaron mejores resultados experimentales que aquellos mas distantes, como español e inglés, de origen latino y germánico respectivamente.

Un análisis usando el *Mann-Whitney-Wilcoxon* test permitió establecer que la distribución de los valores de MOS del algoritmo propuesto aplicado a la traducción voz a voz entre español y catalán es diferente a la correspondiente al algoritmo base (sin hacer transferencia de entonación) con $p < 8\,\%$. En consecuencia, se pudo observar una tendencia en lo referente a la superioridad del algoritmo propuesto sobre el algoritmo base, hecho que también se vió reflejado en los resultados objetivos.

La transferencia del ritmo y la sincronización audio/video fue tratado en la Sección 5.4. Allí se estudiaron diversos algoritmos relacionados con el área de la compresión de audio, con el objeto de su utilización en la transferencia de ritmo y sincronización.

El uso de la información sobre alineamiento, conjuntamente con técnicas de reducción de pausas y silencios, y métodos de compresión del habla a través de la manipulación del ritmo, permitieron obtener resultados experimentales que presentan un retardo medio de 120 milisegundos entre hablante fuente y destino.

La importancia de esta tarea se refleja en la necesidad que los múltiples canales de

comunicación, tales como la voz y los gestos, estén sincronizados para que no existan inconsistencias entre ellos, que dificulten la comprensión por parte de la audiencia.

Finalmente, en la Sección 5.5 se estudió el problema de la transferencia de pausas entre idiomas. Un aspecto importante de este enfoque es que contribuye a preservar las pausas del hablante del idioma origen, y en consecuencia, el significado de la oración, evitando posibles ambigüedades o malinterpretaciones.

El análisis de los datos de entrenamiento permitió concluir que la utilización de la tupla (una unidad de traducción) facilita el proceso de transferencia, ya que pocas pausas ocurren dentro de una tupla. Otro aspecto importante del uso de la tupla es evitar los efectos de la diferencia de orden de las palabras entre diferentes idiomas.

La utilidad de la transferencia de las pausas se reflejó en experimentos que involucraron a los idiomas español e inglés, y la mejora introducida por la propuesta es significativa. Un análisis manual de las pausas transferidas usando tuplas que no se encuentran en el hablante destino (falsos positivos) reveló que un 83 % de ellas se encuentran en la ubicación correcta, debido a que el locutor utilizó una pausa en el idioma origen y decidió no hacerla en el idioma destino.

Si se consideran que todas las pausas transferidas son correctas, lo cual es erróneo en el 17 % de los casos, la F-measure para los cuatro locutores utilizados en los experimentos se sitúa entre 75 % y 80 %, y la precisión se encuentra entre 75 % y 79 %. Estos resultados son superiores a los valores de precisión y F-measure alcanzados por los algoritmos que no usan transferencia de pausas, 58 % a 61 %, y 61 % a 63 %, respectivamente.

Direcciones futuras

Durante el desarrollo de esta tesis han surgido todo un conjunto de aspectos que deben ser estudiados en el futuro. Lo mismo será tarea del autor de esta tesis y de aquellas personas interesadas en las direcciones futuras propuestas.

Por ejemplo, los contornos predichos con JEMA presentan un cierto suavizado que contribuye a percibir la voz sintetizada como menos expresiva, debido a una disminución del rango de frecuencia fundamental. Es de interés estudiar técnicas que permitan evitar este efectos, tales como la eliminación de contornos espúreos en el proceso de estimación de los parámetros.

Otro importante aspecto a investigar es la inclusión de medidas psicoacústicas, tales como la *Just Noticeable Difference* (JND) en la medida de error para la estimación de los parámetros. De esta manera se evitaría considerar como error aquellas diferencias imperceptibles, y se podría hacer énfasis en aquellas más significativas.

En lo referente al modelado de las duraciones, se propusieron algoritmos que utilizan la duración silábica predicha para estimar la duración segmental. Resulta interesante continuar la investigación en esta dirección incluyendo unidades de mayor duración, tales como palabras, grupos acentuales y frases entonativas, para modelar la influencia en el ritmo de cada uno de estos niveles.

Las experiencias en el modelado de las junturas terminales utilizando diferentes modelos estadísticos y unidades (palabras y grupos acentuales) ha revelado pocos progresos, donde las falencias existentes tanto en la estimación de la presencia de junturas como en la evaluación objetiva continúan.

Las limitaciones en el entendimiento del lenguaje natural por parte de los ordenadores constituyen una limitación que impide el progreso en este área. En consecuencia, sería conveniente continuar la investigación en lo relacionado con la evaluación del rendimiento de los modelos, incorporando técnicas aplicadas en la evaluación de la traducción automática.

La técnica utilizada para el modelado de la prosodia es otro punto que se debe revisar. Por ejemplo, técnicas como el razonamiento basado en caso, o por su denominación en inglés *Case Based Reasoning* (CBR), permitirían el modelado usando la analogía, tratando de resolver un problema objetivo a partir de la experiencia acumulada. En la tesis de Iriondo [Iri08] se utiliza esta técnica para el modelado de habla expresiva con buenos resultados, aunque no por igual para todos los estilos.

Los algoritmos para la mejora de la expresividad propuestos en el Capítulo 5 tienen varios aspectos que pueden ser investigados. En esta tesis se propone la codificación de la entonación del hablante origen en forma totalmente automática. Sin embargo, los patrones utilizados para la codificación no poseen hasta el momento una interpretación. Son solamente patrones regulares que permiten mejorar la predicción del contorno de frecuencia fundamental del idioma destino. En consecuencia, es de interés obtener una interpretación de la función o significado de cada patrón de entonación, para tener un mejor entendimiento del fenómeno de la entonación. Como resultado de ello, se podría investigar sobre como obtener un método que permita modelar la multiplicidad de formas de contornos entonativos en ambos idiomas con la misma función.

Se ha propuesto un algoritmo de agrupamiento que usa los contornos de frecuencia fundamental originales de ambos idiomas para la mejora del modelado de la entonación del idioma destino. Una posible dirección futura sería aplicar dicho algoritmo al residuo del contorno predicho por un modelo de entonación (por ejemplo, S-Bezier JEMA) con el objeto de reducir la influencia de la componente de grupo entonativo y resaltar aquellos grupos acentuales que tienen un alto error de predicción con un modelo de entonación aislado.

La aportación en la transferencia de la prosodia realizada en la tesis de Iriondo [Iri08] podría también indicar un camino a explorar. Allí se utilizaron muestras de habla emocionada en castellano (cuatro emociones que tienen una expresión más universal), de las que se extrajeron los valores de los parámetros prosódicos que permitiría modificar el sistema TTS en catalán para generar habla emocionada en esta lengua.

Tal como observa en su artículo Sridhar et al [Sri11], el uso de una anotación más enriquecida para la síntesis aportada por los módulos anteriores ASR y MT puede ser prosódica, como es el enfoque propuesto por esta tesis, o bien siguiendo otras direcciones. Por ejemplo, son ejemplos de tales direcciones el enfoque de traducción voz a voz basado en conceptos descripto por Gu et al [Gu06], o el propuesto por Sridhar et al [Sri11] mediante etiquetas de actos de diálogo y prominencia prosódica.

Apéndice A - Ogmios: el conversor texto a voz de la UPC

Ogmios es el conversor texto a voz multilingüe desarrollado en la Universitat Politècnica de Catalunya [Bon06, Bon07]. El mismo esta compuesto por una serie de algoritmos que son en gran medida independientes del idioma. Es decir, muchos de los algoritmos del sistema pueden ser entrenados para ajustarse a las características de una lengua específica. En consecuencia, mediante el enfoque de aprendizaje automático basado en datos usando un corpus específico del idioma, las particularidades de una lengua pueden ser incorporadas a Ogmios.

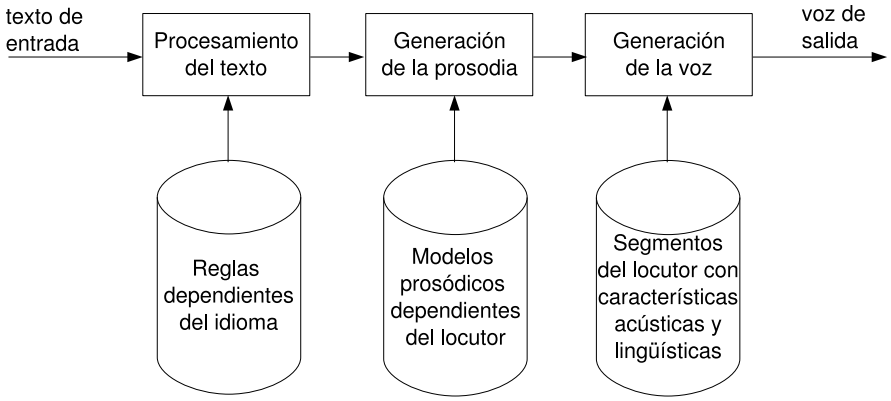


Figura A.1: Diagrama en bloques de los componentes del conversor texto a voz de la UPC: Ogmios.

El sistema fue originalmente desarrollado para el español y el catalán, para luego ser extendido a otros idiomas, tales como inglés, portugués, francés y chino.

El núcleo del sistema es un conjunto de módulos escritos en lenguaje C y C++ con una interfaz común basada en una estructura de datos que describe relaciones lingüísticas entre diferentes niveles, que abarcan desde una descripción básica de la sintaxis del texto, hasta el conjunto de segmentos de habla que serán concatenados en el proceso de síntesis y sus características acústicas. En Ogmios existen varios módulos que pueden interpretar diversos estándares de formato para el texto de entrada, tales como SAPL4, SAPL5, web servers, SABLE, etc.

Ogmios contiene muchos módulos dedicados a diversas tareas específicas. Los mismos

pueden ser agrupados en tres áreas principales: análisis del texto, generación de la prosodia, y generación de la voz, y se describen en forma breve a continuación:

- **Análisis del texto**. En primer lugar, el sistema divide el texto de entrada (texto puro con etiquetas SSML opcionales) en unidades lexicográficas más pequeñas denominadas *tokens*, y clasifica cada una de ellas asignándoles alguna de las diversas categorías utilizadas en la aplicación: puntuación, acrónimo, abreviación, números cardinal, número ordinal, expresión de fecha o números, dirección de internet, etc. Luego, cada *token* puede ser expandido a su forma ortográfica completa, para luego ser etiquetado con marcas morfosintácticas usando un clasificador estadístico. Luego, la pronunciación de cada palabra se obtiene usando diccionarios o reglas. El sistema utiliza transductores de estados finitos para establecer la transcripción fonética de las palabras que no se encuentran en el diccionario, o que no pueden ser predichas por las reglas existentes.
- **Generación de la prosodia**. Este componente es de gran importancia para obtener una calidad de voz natural en la síntesis, e involucra varias tareas, tales como predicción de junturas terminales, entonación, duración segmental, e intensidad. Cada una de ellas es realizada por módulos individuales.
- **Generación de la voz**. En Ogmios la síntesis utiliza el enfoque de concatenación de unidades seleccionadas de segmentos pregrabados que se encuentran en una gran base de datos. Las unidades básicas son los semifonemas dependientes del contexto, y son seleccionadas mediante un algoritmo de programación dinámica basándose en sus características acústicas y fonológicas.

A.1. Procesamiento del texto

La primera tarea del conversor texto a voz es la detección de la estructura del documento y la transformación del texto en palabras. Esta tarea de normalización se realiza utilizando herramientas dependientes del idioma. A través de expresiones regulares de análisis y transformación se convierten los *tokens* en palabras. Las primeras reglas fueron escritas para el español y el catalán, para ser luego extendidas a otros idiomas, como el inglés.

Una vez obtenida una representación del texto de entrada en palabras, un etiquetador morfosintáctico coloca etiquetas en cada una de ellas para describir su función en la oración. Esta información es de gran importancia para uno de los módulos subsiguientes, el conversor de grafemas a fonemas, ya que se utiliza para eliminar ambigüedades en la transcripción fonética.

El conversor de grafemas a fonemas posee varios algoritmos dependiendo del idioma. En el caso del español y el catalán se usan un conjunto de reglas para decidir la transcripción fonética correspondiente a cada palabra. En cambio, para el inglés se usa un diccionario para obtener la transcripción fonética de cada palabra. Si la palabra no existe en el diccionario, se utiliza un transductor de estados finitos para dicha tarea.

Además, el conversor texto a voz de la UPC ofrece la posibilidad de aplicar un conjunto de reglas fonotácticas escritas manualmente luego del proceso de transcripción fonética, con el objeto de introducir diferentes fenómenos encontrados en el habla natural continua: plosivas aspiradas, asimilación de consonantes y la elisión [Agü09].

A.2. Generación de la prosodia

El proceso de generación de la prosodia se puede descomponer en varias tareas, las cuales son realizadas en forma secuencial por diferentes módulos: predicción de junturas terminales, estimación de la duración segmental, predicción de la intensidad y generación del contorno de entonación.

Predicción de junturas terminales

En Ogmios, la predicción de junturas terminales se puede realizar mediante tres algoritmos diferentes. Cada uno de ellos utiliza diversas características extraídas del texto y algoritmos de clasificación entrenados mediante el enfoque de aprendizaje automático basado en datos. Estos algoritmos fueron descritos en la Sección 3.3 de esta tesis.

El primero de ellos modela las junturas terminales usando un árbol de clasificación. Dentro de las características utilizadas se encuentra la ubicación de la última juntura terminal, con el objeto de evitar la predicción de junturas terminales en palabras adyacentes.

Otro de los métodos de predicción de junturas terminales utiliza una combinación de árboles de clasificación y regresión, y un modelo de lenguaje de la secuencia de junturas terminales predichas.

Finalmente, el tercer modelo de junturas terminales hace uso de transductores de estados finitos, que utiliza como información de entrada las etiquetas morfosintácticas y la puntuación del texto.

Una vez predichas las junturas terminales que tienen pausas asociadas, la duración de las pausas se estima usando un árbol de clasificación y regresión.

Estimación de la duración segmental y la intensidad

La duración de los fonemas depende en gran medida de la estructura rítmica del idioma. Por ejemplo, muchos autores indican que el inglés posee isocronía acentual, mientras que el español tiene isocronía silábica.

Ogmios tiene varios algoritmos para la predicción de la duración segmental, y cada uno de ellos tiene en cuenta distintas propuestas de la literatura. Uno de ellos predice la duración de los fonemas de manera individual, sin considerar el nivel suprasegmental. Siguiendo ese enfoque, se encuentran implementados dos algoritmos que utilizan los enfoques CART y suma de productos.

Otro de los algoritmos para la predicción de la duración segmental utiliza la duración

suprasegmental (isocronía silábica o acentual), para luego distribuirla en los segmentos que la constituyen.

La intensidad de los fonemas se predice utilizando CART, usando características similares a las utilizadas para la predicción de la duración segmental.

Generación del contorno de entonación

En Ogmios la entonación se genera a través de dos modelos de entonación: un enfoque superposicional que utiliza JEMA, y otro algoritmo que utiliza selección de unidades.

En el modelo superposicional se combinan los efectos de dos unidades prosódicas: el grupo acentual, que modela los movimientos tonales a nivel de la sílaba acentuada, y los grupos entonativos, que modelan los fenómenos amplios del contorno de entonación. Cada componente del contorno de entonación se aproxima utilizando una curva de Bèzier, usando JEMA para entrenar el modelo.

El segundo modelo utiliza la selección de unidades, ya que se observa en el caso del modelo superposicional una tendencia hacia el suavizado de los contornos resultantes, y la consecuente merma de expresividad de la voz sintética. Los contornos más adecuados se obtienen a través de un agrupamiento utilizando árboles de clasificación, para luego ser seleccionados mediante un algoritmo de programación dinámica para reducir las discontinuidades.

A.3. Generación de la voz

La generación de la voz utiliza el algoritmo Viterbi para la selección de unidades con el objeto de encontrar la secuencia de unidades $u_1 \dots u_n$ del conjunto de la base de datos que minimiza una función de coste con respecto a los valores objetivo $t_1 \dots t_n$. La función está compuesta de un costo objetivo y otro de concatenación, y cada uno es calculado como una suma ponderada de sub-costos individuales:

$$C(t_1 \dots t_n, u_1 \dots u_n) = w^t \sum_{i=1}^{n} \left(\sum_{m=1}^{M^t} w_m^t C_m^t(t_i, u_i) \right) + w^c \sum_{i=1}^{n-1} \left(\sum_{m=1}^{M^c} w_m^c C_m^c(u_i, u_{i+1}) \right) \quad (1)$$

donde w^t y w^c son los pesos globales de los costos objetivo y de concatenación ($w^t + w^c = 1$); M^t es el número de subcostos objetivo y M^c es el número de subcostos de concatenación; $C_m^t(.)$ es el emésimo subcosto objetivo ponderado por el peso w_m^t; mientras que $C_m^c(.)$ es el emésimo subcosto de concatenación ponderado por w_m^c.

Las tablas A.1 y A.2 muestran las características que definen a las funciones de subcostos. Existen dos tipos de funciones de subcosto: binarias, que pueden tener solamente valores 0 o 1, y las continuas, que utilizan una función de distancia sigmoide para acotar su rango entre 0 y 1.

Para ajustar los pesos de manera objetiva se utilizó un enfoque similar al propuesto por Hunt et al. [Hun96]. Para cada par de unidades se computa su distancia usando un vector de características (MFCC, F_0, energía) calculado cada 5ms. Sea $\overline{d}$ el vector de todas las distancias para cada par de unidades, C una matriz donde el elemento $C(i,j)$ es el subcosto j para cada par de unidades i, y $\overline{w}$ el vector de todos los pesos a ser estimados. Si se asume que $C\overline{w} = \overline{d}$, entonces es posible calcular $\overline{w}$ usando regresión lineal. En otras palabras, la función de costo objetivo se convierte en una estimación lineal de la distancia acústica. No existe ajuste automático para los costos de concatenación, y la tarea debe realizarse manualmente.

acento fonético	B
diferencia en la duración	C
diferencia en la energía	C
diferencia en el tono	C
diferencia en el tono al final	C
diferencia de la derivada del tono	C
signo de la derivada del pitch es diferente	B
posición en el grupo acentual	B
trifonema	B
palabra	B

Tabla A.1: Costos objetivos, donde B corresponde a valores binarios y C a valores continuos.

energía	C
tono	C
tono al final	C
distancia espectral	C
concatenación sonora a sorda	B

Tabla A.2: Costos de concatenación, donde B corresponde a valores binarios y C a valores continuos.

En lo relacionado al proceso de generación de la forma de onda, los oyentes otorgan calificaciones más altas a las oraciones sintéticas donde las modificaciones prosódicas son mínimas. Por lo tanto, la mayoría de las unidades seleccionadas son concatenadas usando la información sobre el instante de cierre de la glotis, sin realizar ninguna manipulación prosódica.

A.4. Construcción de la voz sintética

Luego de la normalización y la transcripción fonética de las oraciones, Ogmios es capaz de construir una nueva voz de manera automática partiendo de los ficheros de audio y de la transcripción ortográfica de los mismos.

Este proceso automático consiste en cuatro pasos principales: segmentación automática

de la base de datos, entrenamiento de los modelos prosódicos, ajuste de los pesos de selección, e indexado de la base de datos.

La base de datos es segmentada automáticamente en fonemas usando un alineamiento basado en Modelos Ocultos de Markov, que forma parte de Ramses [Bon98].

Luego, se entrenan modelos HMM de semifonemas dependientes del contexto, para determinar tanto las fronteras de los fonemas, como su punto central. Un modelo de silencio opcional se coloca luego de cada palabra para detectar pausas. Dicho modelo es entrenado en los silencios existentes en los signos de puntuación.

Experimentos previos demuestran que si la transcripción fonética es correcta, los HMM pueden alcanzar una calidad de segmentación similar a la manual [Mak00, Ade05]. Por lo tanto, se presta especial atención a la transcripción fonética y a la eliminación de unidades segmentadas en forma incorrecta.

Debido a que la transcripción fonética automática de una base de datos para síntesis de voz considera las variantes de pronunciación, los errores de articulación y el ruido de grabación, se utilizan varias transcripciones para algunas palabras.

Sin embargo, es posible que algunas unidades estén incorrectamente segmentadas en la base de datos, y resulta deseable su eliminación para evitar que sean elegidas en el proceso de selección de unidades. Para ello, se eliminan el 10 % de aquellas unidades con una baja probabilidad de reconocimiento [Ade06].

Apéndice B - Herramientas estadísticas utilizadas

B.1. Error cuadrático medio

El error cuadrático medio es la métrica dominante en el análisis cuantitativo de rendimiento en el campo del procesamiento de señales. Entre sus aplicaciones se encuentran: es un criterio estándar para el estudio de la calidad y la fidelidad de la señales, es un método útil para la comparación y selección de algoritmos de procesamiento de señales, y es ampliamente utilizado para la optimización en los algoritmos de procesamiento de señales.

El error cuadrático medio se define para señales discretas usando operaciones matemáticas simples. Suponga que x e y son dos señales discretas con una cantidad limitada de muestras N, y que x_i e y_i son los valores de la muestra i-ésima de x e y. El error cuadrático medio (MSE: Mean Squared Error) entre estas dos señales es:

$$MSE(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2$$

En analogía a la desviación estándar, tomando la raiz cuadrada del MSE obtenemos el RMSE (*root mean squared error*), el cual tiene las mismas unidades de la cantidad que está siendo estimada.

Su amplia aplicación surge de diferentes características que lo distinguen de otras medidas de distancia:

- **Cálculo simple**. El cálculo del error cuadrático medio solamente necesita de una multiplicación y dos sumas por cada muestra. Además, es una medida que no posee memoria, y puede ser evaluada para cada muestra en forma independiente de las otras.
- **Métrica de distancia de los espacios euclidianos N-dimensionales**. Todas las normas l_p son métricas de distancia validas en $\mathbb{R}^N$ que satisfacen condiciones que son convenientes para una interpretación directa de similitud: no negatividad ($d_p(x, y) \geq 0$), identidad ($d_p(x, y) = 0$, si y solo si $x = y$), simetría ($d_p(x, y) = d_p(y, x)$) y desigualdad triangular ($d_p(x, z) \leq d_p(x, y) + d_p(y, z)$). El caso $p = 2$ (el RMSE) es la distancia utilizada en los espacios euclidianos N-dimensionales.

- **Significado físico**. El error cuadrático medio tiene un significado físico como una medida de la energía de la señal de error. Tal energia se preserva luego de cualquier transformación lineal ortogonal (o unitaria), como es el caso de la transformada de Fourier. Esta propiedad distingue d_2 de las otras medidas de energia d_p, las cuales no preservan la energía en el campo transformado.

- **Optimización**. El MSE es una métrica excelente para los problemas de optimización por sus propiedades de convexidad, simetría y diferenciabilidad. Muchos problemas de optimización de mínimo MSE (MMSE) tienen una solución analítica cerrada. Cuando ello no es posible, los procedimientos numéricos iterativos son fáciles de formular, ya que el gradiente y la matriz Hessiana son fáciles de calcular.

- **Medida estadística**. El MSE es una medida muy utilizada en los campos de la estadística y la estimación. El error cuadrático medio es el segundo momento del error, y por lo tanto, mide tanto la varianza del estimador como su sezgo: $MSE(\hat{\theta}) = E\left[(\theta - \hat{\theta})^2\right]$. Un MSE igual a cero significa que el estimator $\hat{\theta}$ predice las observaciones del parámetro θ con una exactitud perfecta, que es el objetivo y forma la base para el análisis por regresión usando la minimización del error cuadrático medio.

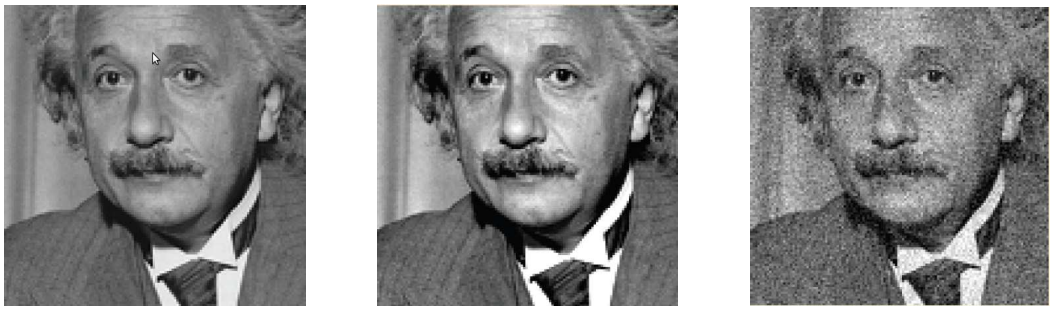
Entre las desventajas del error cuadrático medio, tal como ocurre con la varianza, es que otorga un mayor peso a los *outliers*. Esto es el resultado de elevar al cuadrado cada término, lo cual tiende a dar un peso mayor a los errores grandes que a los pequeños. Esta propiedad no es deseable en algunas aplicaciones, y ha llevado a algunos investigadores a la utilización de algunas alternativas, tales como el error medio absoluto o la mediana.

Otras desventaja es que el error cuadrático medio es independiente de las relaciones temporales o espaciales entre las muestras de la señal original. Esta propiedad no siempre es deseable, ya que estudios sensoriales demuestran que altos valores de MSE no necesariamente implican una diferencia perceptible por un humano [Wan09]. Por ejemplo, en la Figura B.1 se puede observar que imagenes tales como la primera y la segunda, que se perciben como similares, tienen un MSE similar en la comparación de la primera y la última, que se perciben claramente como diferentes.

B.2. Coeficiente de correlación Pearson

En la teoría de probabilidad y la estadística, la correlación (en algunas ocasiones medida como un coeficiente de correlación) indica el grado y la dirección de la relación lineal entre dos variables aleatorias. Esta definición contrasta con el uso coloquial del término, que a veces puede significar una relación no necesariamente lineal.

El coeficiente más conocido es el coeficiente de correlación Pearson, el cual se obtiene dividiendo la covarianza de las dos variables aleatorias X y Y por el producto de sus desviaciones estándar: $\rho = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}$. A pesar de su nombre, el primero que lo introdujo fue Francis Galton [Rod88].



MSE=0 MSE=306 MSE=309

Figura B.1: Comparación de MSE entre diferentes imagenes del físico Eistein. De izquierda a derecha: original, disminución del contraste medio y contaminación con ruido gaussiano. [Imágenes extraídas del artículo de Wang(2009)]

El valor absoluto del coeficiente de correlación Pearson es menor o igual a 1. Las correlaciones iguales a 1 o -1 corresponden a datos cuyos puntos yacen exactamente en una linea. El signo es positivo solamente si la pendiente de la recta de los datos X e Y es de signo positivo. En caso contrario, el signo de la correlación es negativo. En la Figura B.2 se puede observar valores del coeficiente para diferentes distribuciones de puntos.

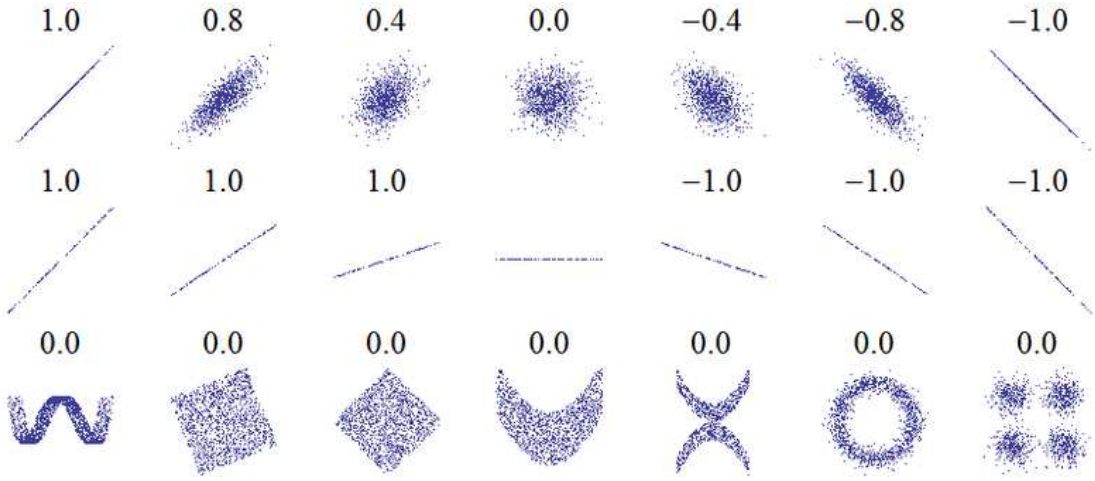


Figura B.2: Valores del coeficiente de correlacion Pearson para diferentes distribuciones de puntos.

Una propiedad matemática importante de este coeficiente es su carácter invariante ante cambios en la ubicación y la escala de los puntos evaluados. En caso de transformar X ($X' = a + bX$) e Y ($Y' = c + dY$), donde a, b, c y d son constantes, el coeficiente de correlación Pearson de X' e Y' será el mismo que para X e Y.

Tal como ocurre con muchas medidas estadísticas, el coeficiente de correlación Pearson puede presentar errores debido a la existencia de observaciones muy diferentes al resto de los datos (*outliers*). En estos casos es necesario el uso de gráficos de dispersión para revelar la existencia de *outliers*.

Otro aspecto a tener en cuenta es que la utilización de este coeficiente de correlación en test estadísticos es dependiente de la distribución de los datos, tal como ocurre con la Transformación de Fisher. Dicha transformación solamente puede ser aplicada si los datos poseen una distribución aproximadamente normal.

Muchos autores ofrecen formas de interpretación para el valor del coeficiente de correlación. Dichos criterios son en cierta medida arbitrarios, y no deben ser utilizados de manera muy estricta. La interpretación del coeficiente depende del contexto y del propósito. Una correlación de 0,9 puede ser baja si se verifica una ley física con instrumentos de alta calidad. Sin embargo, dicho valor puede ser considerado alto en ciencias sociales donde pueden contribuir muchas factores que dificultan la medición.

B.3. Box-plots

Un *box-plot* (también conocido como un diagrama *box-and-whisker*) es una manera conveniente de representar y comparar gráficamente grupos de datos numéricos usando cinco números que resumen su distribución: la observación más pequeña, el primer cuartil (Q1), la mediana (Q2), el tercer cuartil (Q3), y la observación más grande. En un *box-plot* también puede estar indicado si alguna de las observaciones es considerada *outlier*. El *boxplot* fue inventado en 1977 por el estadista estadounidense John Tukey.

Los *box-plots* son una manera útil de mostrar diferencias entre poblaciones sin hacer suposiciones acerca de la distribución estadística subyacente. El espaciado entre las diferentes partes de la caja permite tanto indicar el grado de dispersión y el sesgado en los datos, así como también identificar *outliers*.

Dado un conjunto de datos, un *box-plot* puede ser construido siguiendo los siguientes pasos:

- Se calcula el primer cuartil, la mediana y el tercer cuartil: $Q1$, $Q2$ y $Q3$.
- Luego, se calcula el rango entre cuartiles (IQR: Inter Quartile Range) restando el primer cuartil del tercero ($IQR = Q3 - Q1$).
- Se construye una caja que abarca desde el primer cuartil hasta el tercero.
- Se indica la mediana mediante un símbolo o una linea que divide la caja en el valor de la mediana.
- El valor medio de los datos también es indicado con un punto.
- Cualquier observación menor que $Q1 - 1{,}5IQR$ o superior a $Q3 + 1{,}5IQR$ se considera un *outlier*, y se grafica con un círculo. La manera de indicar el valor más pequeño que no es un *outlier* es a través de una linea que lo conecta con la caja. Lo mismo se realiza con el valor más grande.
- Cualquier observación menor que $Q1 - 3IQR$ o superior a $Q3 + 3IQR$ se considera un *outlier* extremo, y se grafica con un punto.

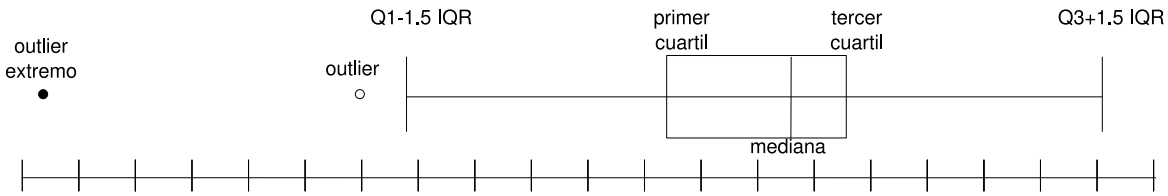


Figura B.3: Ejemplo de un *box-plot*.

B.4. Wilcoxon test

El test de Wilcoxon, o también llamado test de Mann-Whitney-Wilcoxon, es un test no paramétrico para determinar si dos conjuntos de observaciones provienen de la misma distribución. Es uno de los test más conocidos de significancia no paramétricos.

El test fue propuesto inicialmente por Wilcoxon [Wil45], para observaciones del mismo tamaño, y fue extendido para poblaciones de tamaño arbitrario por Mann y Whitney [Man47].

Sean X e Y dos variables aleatorias con funciones densidad de probabilidad acumulada f y g, Mann y Whitney propusieron un indicador estadístico U para verificar la hipótesis de que $f = g$, que depende del ordenamiento relativo de X e Y.

El test U es útil en las mismas situaciones que el test t-Student. Sin embargo, el test U resulta preferible en situaciones donde los datos son ordinales pero no posee un valor de intervalo fijo. Además presenta una mayor robustez a la presencia de datos espurios que podrían llevar a indicar una significancia incorrecta.

En la utilización de este test se asume que:

- Todas las observaciones de ambos grupos son independientes entre sí.
- Las observaciones corresponden a valores continuos que poseen un orden. Es decir, entre dos observaciones se puede establecer cual de ellas es la mayor.
- La hipótesis nula establece que las distribuciones de ambos grupos es la misma. La probabilidad de una observación de la población X que exceda una observación de la segunda población Y es la misma que una observación de la población Y exceda una observación de X. Es decir, existe una simetría entre las poblaciones con respecto a la probabilidad de extracción aleatoria de una observación más grande.

Apéndice C - Corpus TC-STAR

C.1. Corpus monolingüe

Los experimentos para la generación de prosodia se realizaron usando las voces base del proyecto europeo TC-STAR. Estas bases de datos constan de dos hablantes, siendo uno de cada sexo. El corpus está compuesto de los siguientes escenarios:

- **C1.1** Habla transcrita paralela perteneciente a diferentes dominios.
- **C1.2** Habla general transcrita correspondiente a diferentes dominios.
- **C2** Novelas e historias pequeñas con oraciones cortas (texto escrito de diferentes dominios).
- **C3.1** Frases frecuentes. Sirve para mejorar la calidad de frases usadas usualmente tales como frases que contienen fechas, números, expresiones si/no y frases frecuentes encontradas en los dominios definidos en las especificaciones del proyecto LC-STAR.
- **C3.2** Oraciones para cobertura de trifonemas. Este sirve para mejorar la cobertura de los segmentos del habla con respecto a sílabas o fonemas que son raros de encontrar.
- **C3.3** Oraciones de imitación. Esta porción del corpus tiene como finalidad la investigación de la conversión de voz. El corpus contiene oraciones con una alta cobertura de todos los fonemas del idioma incluyendo los fonemas con poca frecuencia. Las oraciones fueron leídas usando imitación.

Debido a la restricción impuesta de un corpus de 10 horas de duración para un hablante, se decidió enfocar los esfuerzos principalmente en la cobertura fonética y de las variaciones prosódicas. La voz debe sonar como si fuera producida por un traductor competente hablando de una manera bastante neutral.

Las condiciones de grabación fueron:

- Frecuencia de muestreo: 96 Khz.
- Ancho de banda: 40-20000 Khz.

- Precisión: 24 bits.
- $SNR_A > 40dBA$.
- Reverberación: $RT60 < 0,3s$.
- Tres canales: micrófono de membrana, laringógrafo y micrófono cercano.

La anotación y la segmentación de la base de datos se basó en las siguientes reglas:

- Todas las grabaciones son normalizadas, se colocan las etiquetas morfosintácticas y se anotan con marcadores específicos, los cuales son importantes para seleccionar las unidades del habla, tales como ruido, palabras ininteligibles, etc.
- Las grabaciones también son etiquetadas prosódicamente. Las fronteras entonativas se anotan usando dos niveles: grupo y cláusula entonativa. Los acentos marcados con una prominencia entonativa se anotan usando dos niveles: normal y enfático.
- La transcripción fonética se realiza manualmente escuchando a las grabaciones.
- Las grabaciones se dividen completamente en segmentos del habla. Dos horas de la base de datos se revisan manualmente por parte del productor.
- La señal de habla se etiqueta completamente con marcas glotales y dos horas son revisadas manualmente por el productor.

C.2. Corpus bilingüe

Con el objeto de generar voces para la investigación del habla expresiva en español e inglés, dos locutores de sexo femenino y dos de sexo masculino fueron grabados pronunciando 220 párrafos del dominio parlamentario.

Para obtener un estilo de habla adecuado, a los locutores se les reproduce la voz original del parlamentario que pronunció el párrafo que está siendo grabado. Tanto la entonación como la expresión, cadencia y pausas deben ser reproducidos para generar una voz expresiva adecuada.

A fin de lograr una mayor consistencia en los datos, se le solicitó a cada locutor que utilice el mismo estilo para pronunciar el mismo párrafo en cada idioma. Un mismo párrafo fue grabado en cada idioma de manera consecutiva, para permitir al locutor recordar el estilo utilizado.

Cada audio generado contiene un texto asociado con la transcripción ortográfica del párrafo pronunciado por el locutor. No existe transcripción fonética manual, y como resultado de ello, tampoco se posee segmentación manual en fonemas. Tanto la transcripción como la segmentación fonética fueron realizadas de manera automática.

Publicaciones

- **2003**

  Phrase break prediction: a comparative study. Juan Carlos Tulli, Esteban Lucio González and Pablo Daniel Agüero.. XIX Congreso de la Sociedad Española para el procesamiento del Lenguaje Natural. Alcala de Henares, Spain. September, 2003.

- **2004**

  Automatic Analysis and Synthesis of Fujisaki's Intonation Model for TTS. Pablo Daniel Agüero, Klaus Wimmer and Antonio Bonafonte. Speech Prosody 2004. Nara, Japan. March, 2004.

  Intonation Modeling for TTS using a Joint Extraction and Prediction Approach. Pablo Daniel Agüero and Antonio Bonafonte. 5th ISCA Speech Synthesis Workshop. Pittsburgh, USA. June, 2004.

  Intonation Modeling Using a joint extraction and prediction approach. Pablo Daniel Agüero and Antonio Bonafonte. 11th International Workshop .Advances in Speech Technology 2004". Maribor, Slovenia. July, 2004.

  Phrase Break Prediction Using a Finite State Transducer. Antonio Bonafonte and Pablo Daniel Agüero. 11th International Workshop .Advances in Speech Technology 2004". Maribor, Slovenia. July, 2004.

  Joint Extraction and Prediction of Fujisaki's Intonation Model Parameters. Pablo Daniel Agüero, Klaus Wimmer and Antonio Bonafonte. ICSLP 2004. Jeju Island, Korea. October, 2004.

  Els Talps També Parlen. Ignasi Esquerra, Jordi Adell, Pablo Daniel Agüero, Antonio Bonafonte, Helenca Duxans, Asunción Moreno, Javier Pérez and David Sündermann. CELC 2004. Andorra. November, 2004.

- **2005**

  Improving TTS quality using pitch contour information of source speaker in S2ST framework. Pablo Daniel Agüero, Jordi Adell and Antonio Bonafonte. 12th International Workshop .Advances in Speech Technology 2005". Maribor, Slovenia. July, 2005.

  Training the Tilt Intonation Model using the JEMA methodology. Matej Rojc, Pablo Daniel Agüero, Antonio Bonafonte and Zdravko Kacic. Eurospeech 2005. Lisboa, Portugal. September, 2005.

Consistent estimation of Fujisaki's intonation model parameters. Pablo Daniel Agüero and Antonio Bonafonte. SPECOM 2005. Patras, Greece. October, 2005.

- **2006**

  Spanish synthesis corpora. Marti Umbert, Asunción Moreno, Pablo Daniel Agüero and Antonio Bonafonte. LREC 2006. Genoa, Italy. May 24-26, 2006.

  Ogmios: the UPC text-to-speech synthesis system for spoken translation. Antonio Bonafonte, Pablo Daniel Agüero, Jordi Adell and Asuncion Moreno. TC-Star Workshop on Speech-to-Speech Translation. Barcelona, Spain. June 19-21, 2006.

  Database Pruning for unsupervised building of Text-to-Speech voices. Jordi Adell, Pablo D. Agüero and Antonio Bonafonte. International Conference on Audio Speech and Signal Processing , ICASSP. Toulouse, France. May, 2006.

  Prosody Generation for Speech-to-Speech Translation. Pablo Daniel Agüero, Jordi Adell and Antonio Bonafonte. International Conference on Audio Speech and Signal Processing , ICASSP. Toulouse, France. May, 2006.

  Facing data scarcity using variable feature vector dimension. Pablo Daniel Agüero and Antonio Bonafonte. Speech Prosody 2006. Dresden, Germany. May, 2006.

  Prosody generation in the Speech-to-Speech Translation Framework. Pablo Daniel Agüero, Jordi Adell and Antonio Bonafonte. Speech Prosody 2006. Dresden, Germany. May, 2006.

- **2007**

  Intonation model training using the Variable Feature Vector Dimension Approach. Pablo Daniel Agüero, Juan Carlos Tulli and Antonio Bonafonte. XII RPIC. Rio Gallegos, Argentina. October 16-18, 2007.

  The UPC TTS system description for the 2007 Blizzard Challenge. Antonio Bonafonte, Jordi Adell, Pablo D. Agüero, Daniel Erro, Ignasi Esquerra, Asunción Moreno, Javier Pérez and Tatyana Polyakova. 3rd Blizzard Challenge. Bonn, Germany. August 25, 2007.

  Ogmios in the 2007 Evaluation Campaign. Antonio Bonafonte, Asuncion Moreno, Jordi Adell, Pablo D. Agüero, Daniel Erro, Javier Perez, Ignasi Esquerra and Tatyana Polyakova. 2007 TC-Star Workshop. Aachen, Germany. March 28-30, 2007.

- **2008**

  Pause Transfer in the Speech-to-Speech Translation Domain. Pablo Daniel Agüero, Juan Carlos Tulli and Antonio Bonafonte. Speech Prosody 2008. Campinas, Brazil. May 6-9, 2008.

  A New Clustering Approach for JEMA. Pablo Daniel Agüero, Juan Carlos Tulli and Antonio Bonafonte. Speech Prosody 2008. Campinas, Brazil. May 6-9, 2008.

  A Study of JEMA for Intonation Modeling. Pablo Daniel Agüero, Juan Carlos Tulli and Antonio Bonafonte. ICASSP 2008. Las Vegas, USA. March 30-April 4, 2008.

Bibliografía

[Acapel] *http://www.acapela-group.com/text-to-speech-interactive-demo.html.*

[Ada93] S. G. Adams, G. Weismer, and R. D. Kent, "Speaking rate and speech movement velocity profiles", *Journal of Speech and Hearing Research*, Vol. 36, pags. 41–54, 1993.

[Ade05] J. Adell, and A. Bonafonte, "Towards phone segmentation for concatenative speech synthesis", *Proceedings of 5th Speech Synthesis Workshop*, pags. 139–144, 2005.

[Ade06] J. Adell, P.D. Agüero, and A. Bonafonte, "Database pruning for unsupervised building of text-to-speech voice", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 889–892, 2006.

[Agü04a] P. D. Agüero, and A. Bonafonte, "Intonation modeling for TTS using a joint extraction and prediction approach", *Proceedings of the International Workshop on Speech Synthesis*, pags. 67–72, 2004.

[Agü04b] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Automatic analysis and synthesis of Fujisaki's intonation model for TTS", *Proceedings of the International Conference on Speech Prosody*, pags. 427–430, 2004.

[Agü04c] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Joint extraction and prediction of Fujisaki's intonation model parameters", *Proceedings of the International Conference on Spoken Language Processing*, pags. 757–760, 2004.

[Agü05] P.D. Agüero, and A. Bonafonte, "Consistent estimation of Fujisaki's intonation model parameters", *Proceedings of SPECOM*, 2005.

[Agü09] P. D. Agüero, A. Bonafonte, and J. C. Tulli, "Improving consistence of phonetic transcription for text-to-speech", *Proceedings of Interspeech 2009*, pags. 536–539, 2009.

[Alc98] S. Alcoba, and J. Murillo, "Intonation in Spanish", *In S. Young and G. Bloothooft. editors, Intonation Systems. A Survey of Twenty Languages, Cambridge University Press*, pags. 152–167, 1998.

[All87] J. Allen, M. S. Hunnicutt, and D. Klatt, "From text to speech. The MITalk system", *Cambridge: Cambridge University Press.*, 1987.

[Alm97] M. Almeida, "Organización temporal del español: el principio de isocronía", *Revista de Filología Románica*, , nº 14, pags. 29–40, 1997.

[And84] M. Anderson, J. Pierrehumbert, and M. Liberman, "Synthesis by rule of English intonation patterns", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 281–284, 1984.

[Ant03] E. Rodero Antón, "Locución radiofónica", *Publicaciones de la Universidad Pontificia de Salamanca*, 2003.

[Aro97] B. Arons, "Speechskimmer: A system for interactively skimming recorded speech", *ACM Transactions on Computer-Human Interaction*, pags. 3–38, 1997.

[Ata82] B. S. Atal, and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 614–617, 1982.

[ATalke] *http://www.atutor.ca/atalker/*.

[ATTSit] *http://www2.research.att.com/ ttsweb/tts/demo.php.*

[Bai03] G. Bailly, N. Campbell, and B. Mobius, "ISCA Special Session: Hot Topics in Speech Synthesis", *Proceedings of Eurospeech*, pags. 37–40, 2003.

[Bar94] P. Barbosa, and G. Bailly, "Characterisation of rhythmic patterns for text-to-speech synthesis", *Speech Communication*, Vol. 15, pags. 127–137, 1994.

[Bea10] R. Beaufort, S. Roekhaut, L. Cougnon, and C. Fairon, "A hybrid rule/model-based finite-state framework for normalizing sms messages", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pags. 770–779, 2010.

[Bec86] M. Beckman, and J. Pierrehumbert, "Japanese prosodic phrasing and intonation synthesis", *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pags. 173–180, 1986.

[Bes01] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, and D. Vaufreydaz, "Speech translation for french in the nespole! european project", *Proceedings of Eurospeech*, 2001.

[Bis08] M. Bisani, and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", *Speech Communication*, Vol. 50, pags. 434–451, 2008.

[Bla96] A. Black, and A. Hunt, "Generating F0 contours from ToBI labels using linear regression", *Proceedings of the International Conference on Spoken Language Processing*, pags. 1385–1388, 1996.

[Bla97] A. Black, and P. Taylor, "Assigning phrase breaks from part-of-speech sequences", *Proceedings of the European Conference on Speech Communication and Technology*, pags. 995–998, 1997.

[Blo97] H. U. Block, "The language components in Verbmobil", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 79–82, 1997.

[Boe] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer", *http://www.fon.hum.uva.nl/praat/*.

[Bon96] A. Bonafonte, "Language modeling using x-grams", *Proceedings of the International Conference on Spoken Language Processing*, pags. 394–397, 1996.

[Bon98] A. Bonafonte, J. B. Mariño, A. Nogueiras, and J. A. Rodríguez Fonollosa, "RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC", *VIII Jornadas de Telecom I+D (TELECOM I+D'98), Madrid, Spain*, 1998.

[Bon04] A. Bonafonte, and P.D. Agüero, "Phrase break prediction using a finite state transducer", *Proceedings of the 11th International Workshop Advances in Speech Technology*, 2004.

[Bon06] A. Bonafonte, P.D. Agüero, J. Adell, J. Pérez, and A. Moreno, "OGMIOS: The UPC text-to-speech synthesis system for spoken translation", *TC-Star Workshop on Speech to Speech Translation*, pags. 199–204, 2006.

[Bon07] A. Bonafonte, J. Adell, P.D. Agüero, D. Erro, I. Esquerra, A. Moreno, J. Pérez, and T. Polyakova, "The UPC TTS system description for the 2007 Blizzard Challenge", *Proceedings of 6th ISCA Workshop on Speech Synthesis*, pags. 1–4, 2007.

[Bot01] A. Botinis, B. Granström, and B. Möbius, "Developments and paradigms in intonation research", *Intonation. Special issue, Speech Communication*, Vol. 33, pags. 263–296, 2001.

[Bou04] C. Bouzon, and D. Hirst, "Isochrony and prosodic structure in British English", *Proceedings of the International Conference on Speech Prosody*, pags. 223–226, 2004.

[Bou08] P. Bouillon, G. Flores, M. Georgescul, S. Halimi, B. A. Hockey, H. Isahara, K. Kanzaki, Y. Nakao, M. Rayner, M. Santaholma, M. Starlander, and N. Tsourakis, "Many-to-many multilingual medical speech translation on a PDA", *Proceedings of The 8th Conference of the Association for Machine Translation*, pags. 314–323, 2008.

[Bra03] D. Bradley, E. Fernández, and D. Taylor, "Prosodic weight versus information load in the RC attachment ambiguity", *16th Annual CUNY Conference on Human Sentence Processing*, 2003.

[Bra07] T. Brants, A. Popat, P. Xu, F. Och, and J. Dean, "Large language models in machine translation", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pags. 858–867, 2007.

[Bre84] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees", *Chapman & Hall*, 1984.

[Bri95] E. Brill, "Unsupervised learning of disambiguation rules for part of speech tagging", *Proceedings of 3rd Workshop on Very Large Corpora*, pags. 1–13, 1995.

[Bur03] S. Burger, E. Costantini, and F. Pianesi, "The NESPOLE! multimodal speech-to-speech translation system: User based system improvements", *Proceedings of the 8th International Conference on Human Aspects of Advanced Manufacturing*, 2003.

[Bur04] D. C. Burnett, M. R. Walker, and A. Hunt, "Speech Synthesis Markup Language (SSML) Version 1.0", *W3C Recommendation 7 September 2004*, 2004.

[Cab04] M. Caballero, A. Moreno, and A. Nogueiras, "Data driven multidialectal phone set for Spanish dialects", *Proceedings of the International Conference on Spoken Language Processing*, pags. 837–840, 2004.

[Cam91] N. Campbell, and S.D. Isard, "Segment durations in a syllable frame", *Journal of Phonetics*, , nº 19, pags. 29–38, 1991.

[Cam92a] N. Campbell, "Multi-level timing in speech", *PhD thesis*, 1992.

[Cam92b] N. Campbell, "Syllable-based segmental duration", *Talking machines. Theories, models and designs*, pags. 211–224, 1992.

[Cam93] N. Campbell, "Predicting segmental durations for accommodation within a syllable-level timing framework", *Proceedings of Eurospeech*, pags. 1081–1084, 1993.

[Car91] M. Carrió, and A. Ríos, "Compensatory shortening in Spanish spontaneous speech", *Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, 1991.

[Cet99] M. Cettolo, A. Corazza, G. Lazzari, F. Pianesi, E. Pianta, and L. M. Tovena, "A speech-to-speech translation based interface for tourism", *Proceedings of the ENTER Conference*, 1999.

[Cha89] F. Charpentier, and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Proceedings of Eurospeech*, pags. 13–19, 1989.

[Che08] A. Chella, R. E. Barone, G. Pilato, and R. Sorbello, "An emotional storyteller robot", *Proceedings of the AAAI Spring Symposium*, pags. 17–22, 2008.

[Cho01] T. Cho, "Effects of prosody on articulation in English", *Doctoral Dissertation, University of California, Los Angeles, EEUU*, 2001.

[Cis07] Cisco, "Say it smart specifications for cisco unified customer voice portal", Tech. rep., Cisco, 2007.

[Cor01] R. Cordoba, J.M. Montero, J. Gutierrez-Arriola, and J.M. Pardo, "Duration modeling in a restricted-domain female-voice synthesis in Spanish using neural networks", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 793–796, 2001.

[Cov98] M. Covell, M. Withgott, and M. Slaney, "Mach1: nonuniform time-scale modification of speech", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pags. 349–352, 1998.

[Cre04] J. M. Crego, J. B. Mariño, and A. de Gispert, "Finite-state-based and phrase-based statistical machine translation", *Proceedings of the International Conference on Spoken Language Processing*, pags. 37–40, 2004.

[Cór04] R. Córdoba, F. Fernández, V. Sama, L.F. D'Haro, R. San-Segundo, J.M. Montero, J. Macías-Guarasa, J. Ferreiros, and J.M. Pardo, "Realización de sistemas de diálogo en una plataforma compatible con voicexml: Proyecto gemini", *Congreso Sociedad Española para el Procesamiento del Lenguaje Natural*, 2004.

[d'A95] C. d'Alessandro, and P. Mertens, "Automatic pitch contour stylization using a model of tonal perception", *Computer Speech and Language*, Vol. 9, pags. 257–288, 1995.

[Del66] P. Delattre, "A comparison of syllable length conditioning among languages", *International Review of Applied Linguistics*, Vol. 4, pags. 183–198, 1966.

[Don96] R. Donovan, "Trainable speech synthesis", *PhD Thesis*, 1996.

[Dor98] B. J. Dorr, P. W. Jordan, and J. W. Benoit, "A survey of current paradigms in machine translation", *Technical Report*, 1998.

[DTT03] L. A. Hernández Gómez D. T. Toledano, and L. Villarrubia Grande, "Automatic phonetic segmentation", *IEEE Transactions on Speech and Audio Processing*, pags. 617–625, 2003.

[Dus00] K. Dusterhoff, "Synthesizing fundamental frequency using models automatically trained from data", *PhD thesis*, 2000.

[Déc07] D. Déchelotte, H. Schwenk, G. Adda, and J-L Gauvain, "Improved machine translation of speech-to-text outputs", *Proceedings of Interspeech*, pags. 2441–2444, 2007.

[Eid03] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, and T. Mathes, "Recent improvements to the IBM trainable speech synthesis system", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 708–711, 2003.

[Ekl95] R. Eklund, and B. Lyberg, "Inclusion of a prosodic module in spoken language translation systems", *Journal of the Acoustical Society of America*, Vol. 98, nº 5, pags. 2894–2899, 1995.

[EML] "W3C Emotion Incubator Group http://www.w3.org/2005/incubator/emotion/xgr-emotion-20070710/", .

[Esc02a] D. Escudero, "Modelado estadístico de entonación con funciones de Bézier: Aplicaciones a la conversión texto-voz en Español.", *PhD Thesis, Universidad de Valladolid*, 2002.

[Esc02b] D. Escudero, and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 481–484, 2002.

[Feb98] A. Febrer, J. Padrell, and A. Bonafonte, "Modeling phone duration: Application to Catalan TTS", *Proceedings of the International Workshop on Speech Synthesis*, pags. 43–46, 1998.

[Fla72] J. Flanagan, "Speech analysis, synthesis, and perception", *Springer-Verlag, Berlin-Heidelberg-New York.*, 1972.

[Fla73] J. Flanagan, and L. Rabiner, "Speech synthesis", *Dowden, Hutchinson & Ross, Inc., Pennsylvania.*, 1973.

[For98] C. Fordyce, "Prosody prediction for speech synthesis using transformational rule-based learning", *Master of Science Thesis. Boston University, College of Engineering*, 1998.

[For03] M. Forsberg, "Why is speech recognition difficult?", *Technical Report*, 2003.

[Fre94] R. Frederking, and S. Nirenburg, "Three heads are better than one", *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, pags. 95–100, 1994.

[Fre97] R. Frederking, A. Rudnicky, and C. Hogan, "Interactive speech translation in the DIPLOMAT project", *Workshop on Spoken Language Translation at ACL97*, 1997.

[Fre02] R. Frederking, A. W Black, R. Brown, J. Moody, and E. Steinbrecher, "Field testing the tongues speech-to-speech machine translation system", *Proceedings of the International Conference on Language Resources and Evaluation*, pags. 160–164, 2002.

[Fuj84] H. Fujisaki, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *Journal of the Acoustical Society of Japan*, Vol. 5, pags. 233–242, 1984.

[Fuj98] H. Fujisaki, and S. Ohno, "The use of a generative model of f0 contours for multilingual speech synthesis", *Proceedings of the 4th International Conference on Signal Processing*, pags. 714–717, 1998.

[Fuj00a] H. Fujisaki, S. Narusawa, and M. Maruno, "Pre-processing of fundamental frequency contours of speech for automatic parameter extraction", *Proceedings of the International Conference on Signal Processing*, pags. 722–725, 2000.

[Fuj00b] H. Fujisaki, S. Ohno, and S. Narusawa, "Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common Japanese and the standard Chinese", *Proceedings of the 5th Seminar on Speech Production*, pags. 145–148, 2000.

[Fur95] O. Furuse, J. Kawai, H. Lida, S. Akamine, and D. Kim, "Multi-lingual spoken language translation utilizing translation examples", *Proceedings of NLPRS*, pags. 544–549, 1995.

[Füg06] C. Fügen, M. Kolss, M. Paulik, and A. Waibel, "Open domain speech translation: From seminars and speeches to lectures", *TC-STAR Workshop on Speech-to-Speech Translation*, pags. 81–86, 2006.

[Füg07] C. Fügen, and M. Kolss, "The influence of utterance chunking on machine translation performance", *Proceedings of Interspeech*, pags. 2837–2840, 2007.

[Gal01] L. Galescu, and J. F. Allen, "Bi-directional conversion between graphemes and phonemes using a joint n-gram model", *Proceedings of the 4th ISCA workshop on Speech Synthesis*, pags. 103–108, 2001.

[Gal07] M.J.F. Gales, X. Liu, R. Sinha, P.C. Woodland, K. Yu, S. Matsoukas, T. Ng, K. Nguyen, L. Nguyen, J-L Gauvain, L. Lamel, and A. Messaoudi, "Speech recognition system combination for machine translation", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 1277–1280, 2007.

[Gan88] C. K. Gan, and R. W. Donaldson, "Adaptive silence deletion for speech storage and voice mail applications", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, pags. 924–927, 1988.

[Gar96] J.M. Garrido, "Modelling spanish intonation for text-to-speech applications", *PhD Thesis, Universidad Autónoma de Barcelona*, 1996.

[Gar01] J.M. Garrido, "La estructura de las curvas melódicas del español: propuesta de modelización", *Lingüística Española Actual*, , nº 23, pags. 173–209, 2001.

[Gay40] S. Gili Gaya, "La cantidad silábica en la frase", *Castilla*, Vol. 1, pags. 287–298, 1940.

[GE68] F. Goldman-Eisler, "Psycholinguistics: Experiments in spontaneous speech.", *New York: Academic.*, 1968.

[Gil04] J. Gil, and J. Llisterri, "Fonética y fonología del español en españa (1978-2003)", *Lingüística Española Actual*, 2004.

[Gir09] E. Giraudo, and P. Baggia, "Evalita 2009: Loquendo spoken dialog system", Tech. rep., Loquendo, 2009.

[Gis02] A. de Gispert, and J. B. Mariño, "Using x-grams for speech-to-speech translation", *Proceedings of the International Conference on Spoken Language Processing*, pags. 1885–1888, 2002.

[Gu06] L. Gu, Y. Gao, F. Liu, and M. Picheny, "Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation", *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 14, nº 2, pags. 377–392, 2006.

[Har90] J. Hart, R. Collier, and A. Cohen, "A perceptual study of intonation. An experimental approach to speech melody", *Cambridge University Press*, 1990.

[He00] L. He, and A. Gupta, "User benefits of non-linear time compression", *Microsoft Research Technical Report MSR-TR-2000-96, Microsoft*, 2000.

[He01] L. He, and A. Gupta, "Exploring benefits of non-linear timecompression", *Proceedings of the Conference on Multimedia*, pags. 382–391, 2001.

[Hir93] D. Hirst, and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function", *Travaux de l'Institut de Phonétique d'Aixen-Provence*, pags. 75–85, 1993.

[Hir94] D.J. Hirst, N. Ide, and J. Veronis, "Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project", *Proceedings of 2nd ESCA/IEEE Workshop on Intonation*, pags. 77–80, 1994.

[Hir00] D. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems", *Prosody : Theory and Experiment*, 2000.

[Hir03] K. Hirose, Y. Furuyama, S. Narusawa, and N. Minematsu, "Use of linguistic information for automatic extraction of F0 contour generation process model parameters", *Proceedings of Eurospeech*, pags. 141–144, 2003.

[Hor03] C. Hori, and S. Furui, "A new approach to automatic speech summarization", *IEEE Transactions Multimedia*, pags. 368–378, 2003.

[Hua02] J. Huang, and G. Zweig, "Maximum entropy model for punctuation annotation from speech", *Proceedings of the International Conference on Spoken Language Processing*, pags. 917–920, 2002.

[Hun96] A. Hunt, and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 373–376, 1996.

[Hun00] A. Hunt, "JSpeech Markup Language", *W3C Note 05 June 2000*, 2000.

[Iri08] I. Iriondo, "Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva", *PhD thesis*, 2008.

[Jan04] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech", *Speech Communication*, Vol. 42, pags. 155–173, 2004.

[Kai92] N. Kaiki, and Y. Sagisaka, "Pause characteristics and local phrase-dependency structure in Japanese", *Proceedings of the International Conference on Spoken Language Processing*, pags. 357–360, 1992.

[Kar98] M. Karjalainen, T. Altosaar, and M. Vainio, "Speech synthesis using warped linear prediction and neural networks", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 877–880, 1998.

[Kaw99] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds", *Speech Communication*, Vol. 27, pags. 187–207, 1999.

[Kim01] J.-H. Kim, and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition", *Proceedings of Eurospeech*, pags. 2757–2760, 2001.

[Kin10] S. King, "Speech synthesis without the right data", *Proceedings of 7th ISCA Workshop on Speech Synthesis*, pag. 38, 2010.

[Kla76] D.H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *Journal of the Acoustical Society of America*, Vol. 59, nº 5, pags. 1208–1220, 1976.

[Kla87] D.H. Klatt, "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America*, Vol. 82, nº 3, pags. 137–181, 1987.

[Kla01] E. Klabbers, and R. Veldhuis, "Reducing audible spectral discontinuities", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, nº 1, pags. 39–51, 2001.

[Kle98] K. Kleijn, and K. Paliwal, "Speech coding and synthesis", *Elsevier Science B.V., The Netherlands*, 1998.

[Koe00] P. Koehn, S. Abney, J. Hirschberg, and M. Collins, "Improving intonational phrasing with syntactic information", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pags. 1289–1292, 2000.

[Kor97] R. Kortekaas, and A. Kohlrausch, "Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli", *Journal of the Acoustical Society of America*, Vol. 101, pags. 2202–2213, 1997.

[Krö92] B. Kröger, "Minimal rules for articulatory speech synthesis", *Proceedings of EUSIPCO92*, pags. 331–334, 1992.

[Lai94] U. Laine, M. Karjalainen, and T. Altosaar, "Warped linear prediction (WLP) in speech synthesis and audio processing", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 349–352, 1994.

[Lav96] A. Lavie, A. Waibel, L. Levin, D. Gates, M. Gavaldà, T. Zeppenfeld, P. Zhan, and O. Glickman, "Translation of conversational speech with JANUS-II", *Proceedings of the International Conference on Spoken Language Processing*, Vol. 4, pags. 2375–2378, 1996.

[Lav01] A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei, and F. Balducci, "Architecture and design considerations in NESPOLE!: a speech translation system for E-commerce applications", *Proceedings of the First International Conference on Human Language Technology Research*, pags. 1–4, 2001.

[Leh76] I. Lehiste, J. Olive, and L. Streeter, "Role of duration in disambiguating syntactically ambiguous sentences", *Journal of the Acoustical Society of America*, Vol. 60, nº 5, pags. 1199–1202, 1976.

[Lem99] S. Lemmetty, "Review of speech synthesis technology", *Master's Thesis, Helsinki University of Technology*, 1999.

[Lev86] S. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition", *Computer Speech and Language*, Vol. 1, pags. 29–45, 1986.

[LG94] E. Lopez-Gonzalo, and L.A. Hernandez-Gomez, "Data-driven joint F0 and duration modeling in text to speech conversion for Spanish", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 589–592, 1994.

[LG96] E. Lopez-Gonzalo, and J.M. Rodriguez-Garcia, "Statistical methods in data-driven modeling of Spanish prosody for text-to-speech", *Proceedings of the International Conference on Spoken Language Processing*, pags. 1377–1380, 1996.

[LG97] Eduardo Lopez-Gonzalo, Jose M. Rodriguez-Garcia, Luis Hernandez-Gomez, and Juan M. Villar, "Automatic prosodic modeling for speaker and task adaptation in text-to-speech", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 927–930, 1997.

[Lin80] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design", *IEEE Transactions on Communication*, pags. 84–95, 1980.

[Lin90] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory", *Speech Production and Speech Modelling*, 1990.

[Liu03] F. Liu, Y. Gao, L. Gu, and M. Picheny, "Noise robustness in speech to speech translation", *Proceedings of Eurospeech*, pags. 2797–2800, 2003.

[Liu04a] Y. Liu, "Structural event detection for rich transcription of speech", *Ph.D. thesis, Purdue University*, 2004.

[Liu04b] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, B. Peskin, and M. Harper, "The ICSI-SRI-UW metadata extraction system", *Proceedings of the International Conference on Spoken Language Processing*, pags. 577–580, 2004.

[Loquen] *http://www.loquendo.com/en/news/news_emotional_TTS.htm*.

[Lóp93] E. López, "Estudio de técnicas de procesado lingüistico y acústico para sistemas de conversión texto voz en Español basados en concatenación de unidades", *PhD Thesis, E.T.S. de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid*, 1993.

[Mac76] N. MacDonald, "Duration as a syntactic boundary cue in ambiguous sentences", *Proceedings of the IEEE International Conference ASSP*, pags. 565–572, 1976.

[Mac96] M. Macon, and C. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 361–364, 1996.

[Mac97] M. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. George, "A singing voice synthesis system based on sinusoidal modeling", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 361–364, 1997.

[Mag74] D. T. Magill, and C. K. Un, "Speech residual encoding by adaptive delta modulation with hybrid companding", *Proceedings of The National Electronics Conference*, pags. 403–408, 1974.

[Mak00] M. J. Makashay, C. W. Wightman, A. K. Syrdal, and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis", *Proceedings of the International Conference on Spoken Language Processing*, pags. 431–434, 2000.

[Man47] H.B. Mann, and D.R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other", *Annals of Mathematical Statistics*, pags. 50–60, 1947.

[Man83] A. Borzone de Manrique, and A. Signorini, "Segmental duration and rhythm in Spanish", *Journal of Phonetics*, Vol. 11, pags. 117–128, 1983.

[Mar96] R. Marín, L. Aguilar, and D. Casacuberta, "El grupo acentual categorizado como unidad de análisis sintáctico-prosódico", *XII Congreso de Lenguajes Naturales y Lenguajes Formales*, pags. 23–27, 1996.

[McA86] R. McAulay, and T. Quatieri, "Speech analysis-synthesis based on sinusoidal representation", *Proceedings of ASSP*, pags. 744–754, 1986.

[McC43] W.S. McCulloch, and W. Pitts, "A logical calculus of the idea immanent in nervous activity", *Bulletin of Mathematical Biophysics*, , nº 5, pags. 115–133, 1943.

[McK05] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey, "From text to speech summarization", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pags. 997–1000, 2005.

[Med91] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals", *IEEE Transactions on Signal Processing*, pags. 40–48, 1991.

[Mer97] P. Mertens, F. Beaugendre, and C. d'Alessandro, "Comparing approaches to pitch contour stylization for speech synthesis", *Progress in Speech Synthesis*, pags. 347–364, 1997.

[Mix00] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 1281–1284, 2000.

[Mor98] A. Moreno, and J. B. Mariño, "Spanish dialects: phonetic transcription", *Proceedings of the International Conference on Spoken Language Processing*, pags. 189–192, 1998.

[Mou90] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol. 9, pags. 453–467, 1990.

[Möb95] B. Möbius, "Components of a quantitative model of German intonation", *Proceedings of ICPhS*, Vol. 2, pags. 108–115, 1995.

[Möb96] B. Möbius, and J. van Santen, "Modeling segmental duration in German text-to-speech synthesis", *Proceedings of the International Conference on Spoken Language Processing*, pags. 2395–2398, 1996.

[Möh95] G. Möhler, "Rule based generation of fundamental frequency contours for German utterances", *Proceedings of the 2nd 'Speak!' Workshop*, 1995.

[Mül00] A. Müller, H. Zimmermann, and R. Neuneier, "Robust generation of symbolic prosody by a neural classifier based on autoassociators", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 1285–1288, 2000.

[Nar02a] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "Automatic extraction of model parameters from fundamental frequency contours of english utterances", pags. 1725–1728, 2002.

[Nar02b] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech", pags. 509–512, 2002.

[Nav02a] E. Navas, I. Hernaez, and N. Ezeiza, "Assigning phrase breaks using CART's in Basque TTS", *Proceedings of the International Conference on Speech Prosody*, pags. 527–531, 2002.

[Nav02b] E. Navas, I. Hernaez, and J.M. Sanchez, "Basque intonation modelling for text to speech conversion", *Proceedings of the International Conference on Spoken Language Processing*, pags. 2409–2412, 2002.

[Nie97] A. Niemann, E. Nöth, A. Kießling, R. Kompe, and A. Batliner, "Prosodic processing and its use in Verbmobil", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 1–4, 1997.

[Nol67] A. M. Noll, "Cepstrum pitch determination", *Journal of the Acoustical Society of America*, Vol. 41, pags. 293–309, 1967.

[Nol70] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", *Symposium on Computer Processing in Communication*, 1970.

[Nos07] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis", *IEICE Trans. Inf. & Syst.*, pags. 1406–1413, 2007.

[Nuance] *http://www.nuance.com/realspeak/*.

[Och00] F. Josef Och, and H. Ney, "Improved statistical alignment models", *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pags. 440–447, 2000.

[Och03] F. J. Och, and H. Ney, "A systematic comparison of various statistical alignment models", *Computational Linguistics*, Vol. 29, nº 1, pags. 19–51, 2003.

[Oka99] T. Okadome, T. Kaburagi, and M. Honda, "Relations between utterance speed and articulatory movements", *Proceedings of Eurospeech*, pags. 137–140, 1999.

[Oka03] T. Okadome, T. Kaburagi, and M. Honda, "Local speech rate: Relationships between articulation and speech acoustics", *Proceedings of ICPhS*, pags. 3177–3180, 2003.

[Ols72] C. Olsen, "Rhyhtmical patterns and syllable features of the Spanish sense-group", *Rigault y Charbonneau eds.*, pags. 990–996, 1972.

[O'M73] M. H. O'Malley, D. R. Kloker, and D. Dara-Abrams, "Recovering parentheses from spoken algebraic expressions", *IEEE Transactions on Audio*, Vol. 21, pags. 217–220, 1973.

[Pap02] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "Bleu: a method for automatic evaluation of machine translation", *40th Annual meeting of the Association for Computational Linguistics*, pags. 311–318, 2002.

[Pfi96] H.R. Pfitzinger, "Two approaches to speech rate estimation", *Proceedings of the sixth Australian International Conference on Speech Science and Technology*, pags. 421–426, 1996.

[Phi85] M. S. Phillips, "A feature-based time domain pitch tracker", *Journal of the Acoustical Society of America*, Vol. 79, 1985.

[Pie80] J.B. Pierrehumbert, "The phonetics and phonology of English intonation", *PhD Thesis, MIT*, 1980.

[Pit94] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labelling reliability in the ToBI framework", *Proceedings of the third International Conference on Spoken Language Processing*, Vol. 2, pags. 123–126, 1994.

[Pit06] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The ibm expressive text-to-speech synthesis system for american english", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, nº 4, pags. 1099–1108, 2006.

[Piw02] P. Piwek, B. Krenn, M. Schröder, M. Grice, S. Baumann, and H. Pirker, "RRL: A Rich Representation Language for the description of agent behaviour in NECA", *Proceedings of the AAMAS workshop on Embodied conversational agents*, 2002.

[PM89] B. Pompino-Marschall, "On the psychoacoustic nature of the p-center phenomenon", *Journal of Phonetics*, pags. 175–192, 1989.

[Poi80] G. Pointon, "Is Spanish really syllable-timed?", *Journal of Phonetics*, Vol. 8, pags. 293–304, 1980.

[Pri96] P. Prieto, and J. Hirschberg, "Training intonational phrasing rules automatically for English and Spanish text-to-speech", *Speech Communication*, Vol. 18, pags. 281–290, 1996.

[Qui93] A. Quilis, and J. Fernandez, "Tratado de fonología y fonética española", *Gredos, Madrid*, 1993.

[RA99] Lincoln Robert Audi, University of Nebraska (ed.), *The Cambridge Dictionary of Philosophy*, 1999.

[Rah93] M. Rahim, C. Goodyear, B. Kleijn, J. Schroeter, and M. Sondhi, "On the use of neural networks in articulatory speech synthesis", *Journal of the Acoustical Society of America*, Vol. 93, pags. 1109–1121, 1993.

[Rav96] M. Ravishankar, "Efficient algorithms for speech recognition", *Ph.D. Thesis*, 1996.

[Rod88] J. Rodgers, and W. Nicewander, "Thirteen ways to look at the correlation coefficient", *The American Statistician*, pags. 59–66, 1988.

[Roe06] S. Roehling, B. MacDonald, and C. Watson, "Towards expressive speech synthesis in english on a robotic platform", *Proceedings of the 11th International Australasian Conference on Speech Science and Technology*, pags. 130–135, 2006.

[Roj05] M. Rojc, P. D. Agüero, A. Bonafonte, and Z. Kacic, "Training the tilt intonation model using the jema methodology", *Proceedings of Eurospeech 2005*, pags. 3273–3276, 2005.

[Rud95] A. Rudnicky, "Language modeling with limited domain data", *Proceedings of the ARPA Workshop on Spoken Language Technology*, pags. 66–69, 1995.

[San92] J.P.H. van Santen, "Contextual effects on vowel duration", *Speech Communication*, Vol. 11, pags. 513–546, 1992.

[San94] J.P.H. van Santen, "Assignment of segmental duration in text-to-speech synthesis", *Computer, Speech and Language*, Vol. 8, pags. 95–128, 1994.

[San95] E. Sanders, and P. Taylor, "Using statistical models to predict phrase boundaries for speech synthesis", *Proceedings of European Conference on Speech Communication and Technology*, pags. 1811–1814, 1995.

[Sch68] M. R. Schroeder, "Period histogram and product spectrum: new methods for fundamental frequency measurement", *Journal of the Acoustical Society of America*, Vol. 43, pags. 829–834, 1968.

[Sch85] M. R. Schroeder, and B. S. Atal, "Code-Excited Linear Prediction (CELP): high quality speech at very low bit rates", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pags. 937–940, 1985.

[Sch93] M. Schroeder, "A brief history of synthetic speech", *Speech Communication*, Vol. 13, pags. 231–237, 1993.

[Sch03] M. Schroder, and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching", *International Journal of Speech Technology*, , nº 6, pags. 365–377, 2003.

[Sch09] M. Schröder, *Affective Information Processing*, Chap. Expressive Speech Synthesis: Past, Present, and Possible Futures, Springer London, 2009.

[Sec83] B. G. Secrest, and G. R. Doddington, "An integrated pitch tracking algorithm for speech synthesis", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 1352–1355, 1983.

[Sha95] S. Shaiman, S. G. Adams, and M. D. Z. Kimelman, "Timing relationships of the upper lip and jaw across changes in speaking rate", *Journal of Phonetics*, Vol. 23, pags. 119–128, 1995.

[Sie95] M. A. Siegler, "Measuring and compensating for the effects of speech rate in large vocabulary continuous speech recognition", *Master Thesis*, 1995.

[Sil92] K. Silverman, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody", *Proceedings of the International Conference on Spoken Language Processing*, Vol. 2, pags. 867–870, 1992.

[Sil99] Kim E. A. Silverman, and Jerome R. Bellegarda, "Using a sigmoid transformation for improved modeling of phoneme duration", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 385–388, 1999.

[Sil04] S. Silva, and S. Netto, "Closed-form estimation of the amplitude commands in the automatic extraction of the Fujisaki's model", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 621–624, 2004.

[Son97] G. P. Sonntag, T. Portele, and B. Heuft, "Prosody generation with a neural network: Weighting the importance of input parameters", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 931–934, 1997.

[Sos99] J. M. Sosa, "La entonación del Español", *Editorial Cátedra*, 1999.

[Spr98a] R. Sproat, "Multilingual Text-to-Speech Synthesis", *KLUWER academic publishers*, 1998.

[Spr98b] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington, "SABLE: a standard for TTS markup", *Bell Labs-Lucent Technologies and CSTR-University of Edingburgh*, 1998.

[Sri06] V. K. Rangarajan Sridhar, and S. Narayanan, "Analysis of disfluent repetitions in spontaneous speech recognition", *Proceedings of EUSIPCO*, 2006.

[Sri11] "Enriching machine-mediated speech-to-speech translation using contextual information", *Computer Speech & Language*, 2011.

[Sta86] C. Stanfill, and D. Waltz, "Toward memory-based reasoning", *Communications of the ACM*, Vol. 29, pags. 1213–1228, 1986.

[Ste03] J. Stergar, V. Hozjan, and B. Horvat, "Labeling of symbolic prosody breaks for the slovenian language", *International Journal of Speech Technology*, pags. 289–299, 2003.

[Ste10] I. Steiner, Marc Schröder, M. Charfuelan, and A. Klepp, "Symbolic vs. acoustics-based style control for expressive unit selection", *Proceedings of Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*, 2010.

[Sti96] L. Stifelman, "Augmenting real-world objects: A paper-based audio notebook", *Proceedings of CHI*, pags. 199–200, 1996.

[Sty01] Y. Stylianou, and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 837–840, 2001.

[Sum99] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai, "Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach", *Proceedings of MT Summit*, pags. 229–235, 1999.

[Sun01] X. Sun, and T. H. Applebaum, "Intonational phrase break prediction using decision tree and n-gram model", *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech)*, Vol. 1, pags. 537–540, 2001.

[SVNY97] LLC Springer-Verlag New York (ed.), *An Introduction to Text-to-Speech Synthesis*, 1997.

[Syr00] A. Syrdal, and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling", *Proceedings of the Sixth International Conference on Spoken Language Processing*, pags. 235–238, 2000.

[Syr01] A. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody", *Speech Communication*, Vol. 33, nº 1-2, pags. 135–151, 2001.

[Tak98] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX", *Proceedings of the International Conference on Spoken Language Processing*, pags. 2779–2782, 1998.

[Tak99] T. Takezawa, F. Sugaya, A. Yokoo, and S. Yamamoto, "A new evaluation method for speech translation systems and the case study on ATR-MATRIX from Japanese to English", *Proceedings of Machine Translation VII*, pags. 299–307, 1999.

[Tan02] H. Tanaka, S. Nightingale, H. Kashioka, K. Matsumoto, M. Nishiwaki, T. Kumano, and T. Maruyama, "Speech to speech translation system for monologues-data driven approach", *Proceedings of the International Conference on Spoken Language Processing*, pags. 1717–1720, 2002.

[Tay93] P. Taylor, "Automatic recognition of intonation from f0 contours using rise/fall/connection", *Proceedings of Eurospeech*, pags. 789–792, 1993.

[Tay98] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis system", *Third International Workshop on Speech Synthesis*, pags. 147–151, 1998.

[Tay00] P. Taylor, "Analysis and synthesis of intonation using the Tilt model", *Journal of the Acoustical Society of America*, Vol. 107, nº 3, pags. 1697–1714, 2000.

[Tay09] P. Taylor, *Text-To-Speech Synthesis*, 2009.

[TCSTAR] "TCSTAR: Technology and Corpora for Speech to Speech Translation http://www.tc-star.org/", *European Union grant FP6-506738*.

[Tel05] Telecom, "Realspeak telecom software development kit", Tech. rep., Telecom, 2005.

[Tem06] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems", *IV Jornadas en Tecnologia del Habla*, pags. 1–6, 2006.

[Tod05] T. Toda, and K. Tokuda, "Speech parameter generation algorithm considering global variance for hmm-based speech synthesis", *Proceedings of Eurospeech*, pags. 2801–2804, 2005.

[Tok95] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pags. 660–663, 1995.

[Tol88] G. Toledo, "EL ritmo en el español. Estudio fonético con base computacional", *Madrid: Gredos*, 1988.

[Tom22] T. Navarro Tomás, "La cantidad silábica en unos versos de Rubén Darío", *Revista de Filología Española IX*, pags. 1–29, 1922.

[Tra05] E. Luna Traill, A. Vigueras Ávila, and G.E. Baez Pinal, *Diccionario básico de lingüística*, 2005.

[Tuc00] S. Tucker, and S. Whittaker, "Time is of the essence: An evaluation of temporal compression algorithms", *Microsoft Research Technical Report MSR-TR-2000-96, Microsoft*, 2000.

[UPCTTS] *http://gps-tsc.upc.es/veu/soft/demos/tts.php3*.

[Val91] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA techique", *Proceedings of Eurospeech*, pags. 345–348, 1991.

[Val98] J.A. Vallejo, "Mejora de la frecuencia fundamental en la conversión de texto a voz", *PhD Thesis, E.T.S.I de Telecomunicaciones, Universidad Politécnica de Madrid*, 1998.

[Vap79] V. Vapnik, *Estimation of Dependences Based on Empirical Data [in Russian]*, Nauka, Moscow, 1979.

[Vog08] D. Vogiatzis, C. D. Spyropoulos, S. Konstantopoulos, V. Karkaletsis, Z. Kasap, C. Matheson, and O. Deroo, "An affective robot guide to museums", *Proceedings of the Fourth International Workshop on Human-Computer Conversation*, 2008.

[Vér98] J. Véronis, P. Di Cristo, F. Courtois, and C. Chaumette, "A stochastic model of intonation for text-to-speech synthesis", *Speech Communication*, Vol. 26, pags. 233–244, 1998.

[Wah00] Wolfgang Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, 2000.

[Wai96] A. Waibel, *Survey of the State of the Art in Human Language Technology*, 1996.

[Wai03] A. Waibel, A. Badran, A. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna, and J. Zhang, "Speechalator: two-way speech-to-speech translation on a consumer PDA", *Proceedings of the European Conference on Speech Communication and Technology*, 2003.

[Wai08] A. Waibel, and C. Fügen, "Spoken language translation", *IEEE Signal Processing Magazine*, pags. 70–79, 2008.

[Wan07] D. Wang, and S.S. Narayanan, "Robust speech rate estimation for spontaneous speech", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, nº 8, pags. 2190–2201, 2007.

[Wan09] Z. Wang, and A.C. Bovik, "Mean squared error: Love it or leave it?", *IEEE Signal Processing Magazine*, pags. 98–117, 2009.

[Whi02] L. White, "English speech timing: a domain and locus approach", *PhD Thesis, University of Edinburgh*, 2002.

[Wig92] C. Wightman, "Segmental durations in the vicinity of prosodic phrase boundaries", *Journal of the Acoustical Society of America*, Vol. 91, nº 3, pags. 1707–1717, 1992.

[Wig02] C. W. Wightman, "Tobi or not tobi?", *Proceedings of Speech Prosody*, pags. 25–29, 2002.

[Wil45] F. Wilcoxon, "Individual comparisons by ranking methods", *Biometrics Bulletin*, pags. 80–83, 1945.

[Wit93] M. Withgott, and F. Chen, "Computational models of American speech", *Center for the Study of Language and Information*, 1993.

[Wos93] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, and W. Ward, "Recent advances in JANUS: A speech translation system", *Proceedings ARPA Human Language Technology Workshop*, pags. 211–216, 1993.

[Wri97] H. Wright, and P. Taylor, "Modelling intonational structure using Hidden Markov Models", *Proceedings of ESCA Workshop on Intonation*, pags. 333–336, 1997.

[XH92] X-Huang, F. Alleva, H. Hon, M. Hwang, and R. Rosenfeld, "The SPHINX-II speech recognition system: An overview", *CMU Technical Report CMU-CS-92-112*, 1992.

[Yam95] Y. Yamazaki, "Research activities on spontaneous speech translation", *Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems*, pags. 280–283, 1995.

[Yam04] J. Yamagishi, T. Masuko, and T. Kobayashi, "Mllr adaptation for hidden semi-markov model based speech synthesis", *Proceedings of the 8th International Conference on Spoken Language Processing*, pags. 1213–1216, 2004.

[Yam07] J. Yamagishi, and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", *IEICE Trans. Inf. & Syst.*, pags. 533–543, 2007.

[Yam08] J. Yamagishi, Z. Ling, and S. King, "Robustness of hmm-based speech synthesis", *Proceedings of Interspeech*, pags. 581–584, 2008.

[Yos99] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis", *Proceedings of Eurospeech*, pags. 2347–2350, 1999.

[Zen04] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis", *Proceedings of the International Conference on Spoken Language Processing*, pags. 1393–1396, 2004.

[Zen05] H. Zen, and T. Toda, "An overview of nitech hmm-based speech synthesis system for blizzard challenge 2005", *Proceedings of Interspeech*, pags. 93–96, 2005.