Universitat Ramon Llull

TESI DOCTORAL

Títol Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva

Realitzada per Ignacio Iriondo Sanz

en el Centre Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle

i en el Departament Comunicacions i Teoria del Senyal

Dirigida per Dr. Joan Claudi Socoró Carrié
Dr. Joaquim Llisterri Boix

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

A la meva esposa Titina:

La teva fe amb mi, el teu suport incondicional
i la teva dedicació perseverant envers la família i el treball
han estat l'ànima d'aquest treball.

I també per als nostres fills Elena, Clara i Ignasi:

El vostre afecte i somriure a pesar del temps robat
són per a mi un coixí emocional que no té preu.

Resumen

Esta tesis aborda diferentes aspectos relacionados con la síntesis del habla expresiva. Se parte de la experiencia previa en sistemas de conversión de texto en habla del *Grup en Processament Multimodal* (GPMM) de *Enginyeria i Arquitectura La Salle*, con el objetivo de mejorar la capacidad expresiva de este tipo de sistemas. El habla expresiva transmite información paralingüística como, por ejemplo, la emoción del hablante, su estado de ánimo, una determinada intención o aspectos relacionados con el entorno o con su interlocutor. Los dos objetivos principales de la presente tesis consisten, por una parte, en el desarrollo de un corpus oral expresivo y, por otra, en la propuesta de un sistema de modelado y predicción de la prosodia para su utilización en el ámbito de la síntesis expresiva del habla.

En primer lugar, se requiere un corpus oral adecuado para la generación de algunos de los módulos que componen un sistema de síntesis del habla expresiva. La falta de disponibilidad de un recurso de este tipo motivó el desarrollo de un nuevo corpus. A partir del estudio de los procedimientos de obtención de habla emocionada o expresiva y de la experiencia previa del grupo, se plantea el diseño, la grabación, el etiquetado y la validación del nuevo corpus. El principal objetivo consiste en conseguir una elevada calidad de la señal y una cobertura fonética suficiente (segmental y prosódica), sin renunciar a la autenticidad desde el punto de vista de la expresividad oral. El corpus desarrollado tiene una duración de más de cinco horas y contiene cinco estilos expresivos: neutro, alegre, sensual, agresivo y triste. Al tratarse de habla expresiva obtenida mediante la lectura de textos semánticamente relacionados con los estilos definidos, se ha requerido un proceso de validación que garantice que las locuciones que forman el corpus incorporen el contenido expresivo deseado. La evaluación exhaustiva de todos los enunciados del corpus sería excesivamente costosa en un corpus de gran tamaño. Por otro lado, no existe suficiente conocimiento científico para emular completamente la percepción subjetiva mediante técnicas automáticas que permitan una validación exhaustiva y fiable de los corpus orales. En el presente trabajo se ha propuesto un método que supone un avance hacia una solución práctica y eficiente de este problema, mediante la combinación de una evaluación subjetiva con técnicas de identificación automática de la emoción en el habla. El método propuesto se utiliza para llevar a cabo una revisión automática de la expresividad del corpus desarrollado. Finalmente, una prueba subjetiva con oyentes ha permitido validar el correcto funcionamiento de este proceso automático.

En segundo lugar y, sobre la base de los conocimientos actuales, de la experien-

cia adquirida y de los retos que se deseaban abordar, se ha desarrollado un sistema de estimación de la prosodia basado en corpus. Tal sistema se caracteriza por modelar de forma conjunta las funciones lingüística y paralingüística de la prosodia a partir de la extracción automática de atributos prosódicos del texto, que constituyen la entrada de un sistema de aprendizaje automático que predice los rasgos prosódicos modelados previamente. El sistema de modelado prosódico presentado en este trabajo se fundamenta en el razonamiento basado en casos que se trata de una técnica de aprendizaje automático por analogía. Para el ajuste de algunos parámetros del sistema desarrollado y para su evaluación se han utilizado medidas objetivas del error y de la correlación calculadas en las locuciones del conjunto de test. Dado que las medidas objetivas siempre se refieren a casos concretos, no aportan información sobre el grado de aceptación que tendrá el habla sintetizada en los oyentes. Por lo tanto, se han llevado a cabo una serie de pruebas de percepción en las que un conjunto de oyentes ha puntuado un grupo de estímulos en cada estilo. Finalmente, se han analizado los resultados para cada estilo y se han comparado con las medidas objetivas obtenidas, lo que ha permitido extraer algunas conclusiones sobre la relevancia de los rasgos prosódicos en el habla expresiva, así como constatar que los resultados generados por el módulo prosódico han tenido una buena aceptación, aunque se han producido diferencias según el estilo.

PALABRAS CLAVE: Corpus orales, prosodia, síntesis del habla expresiva, conversión de texto en habla, tecnologías del habla

Resum

Aquesta tesi aborda diferents aspectes relacionats amb la síntesi de la parla expressiva. Es parteix de l'experiència prèvia en sistemes de conversió de text a parla del Grup en Processament Multimodal (GPMM) d'Enginyeria i Arquitectura La Salle, amb l'objectiu de millorar la capacitat expressiva d'aquest tipus de sistemes. La parla expressiva transmet informació paralingüística com, per exemple, l'emoció del parlant, el seu estat d'ànim, una determinada intenció o aspectes relacionats amb l'entorn o amb el seu interlocutor. Els dos objectius principals de la present tesi consisteixen, d'una banda, en el desenvolupament d'un corpus oral expressiu i, d'una altra, en la proposta d'un sistema de modelatge i predicció de la prosòdia per a la seva utilització en l'àmbit de la síntesi expressiva del parla.

En primer lloc, es requereix un corpus oral adequat per a la generació d'alguns dels mòduls que componen un sistema de síntesi del parla expressiva. La falta de disponibilitat d'un recurs d'aquest tipus va motivar el desenvolupament d'un nou corpus. A partir de l'estudi dels procediments d'obtenció de parla emocionada o expressiva i de l'experiència prèvia del grup, es planteja el disseny, l'enregistrament, l'etiquetatge i la validació del nou corpus. El principal objectiu consisteix a aconseguir una elevada qualitat del senyal i una cobertura fonètica suficient (segmental i prosòdica), sense renunciar a l'autenticitat des del punt de vista de l'expressivitat oral. El corpus desenvolupat té una durada de més de cinc hores i conté cinc estils expressius: neutre, alegre, sensual, agressiu i trist. En tractar-se de parla expressiva obtinguda mitjançant la lectura de textos semànticament relacionats amb els estils definits, s'ha requerit un procés de validació que garanteixi que les locucions que formen el corpus incorporin el contingut expressiu desitjat. L'avaluació exhaustiva de tots els enunciats del corpus seria excessivament costosa en un corpus de gran grandària. D'altra banda, no existeix suficient coneixement científic per a emular completament la percepció subjectiva mitjançant tècniques automàtiques que permetin una validació exhaustiva i fiable dels corpus orals. En el present treball s'ha proposat un mètode que suposa un avanç cap a una solució pràctica i eficient d'aquest problema, mitjançant la combinació d'una avaluació subjectiva amb tècniques d'identificació automàtica de l'emoció en el parla. El mètode proposat s'utilitza per a portar a terme una revisió automàtica de l'expressivitat del corpus desenvolupat. Finalment, una prova subjectiva ha permès validar el correcte funcionament d'aquest procés automàtic.

En segon lloc i, sobre la base dels coneixements actuals, de l'experiència adquirida i dels reptes que es desitjaven abordar, s'ha desenvolupat un sistema d'estimació de

la prosòdia basat en corpus. Tal sistema es caracteritza per modelar de forma conjunta les funcions lingüística i paralingüística de la prosòdia a partir de l'extracció automàtica d'atributs prosòdics del text, que constitueixen l'entrada d'un sistema d'aprenentatge automàtic que prediu els trets prosòdics modelats prèviament. El sistema de modelatge prosòdic presentat en aquest treball es fonamenta en el raonament basat en casos, que es tracta d'una tècnica d'aprenentatge automàtic per analogia. Per a l'ajustament d'alguns paràmetres del sistema desenvolupat i per a la seva avaluació s'han utilitzat mesures objectives de l'error i de la correlació calculades en les locucions del conjunt de prova. Atès que les mesures objectives sempre es refereixen a casos concrets, no aporten informació sobre el grau d'acceptació que tindrà la parla sintetitzada en els oïdors. Per tant, s'han portat a terme una sèrie de proves de percepció en les quals un conjunt d'avaluadors ha puntuat un grup d'estímuls en cada estil. Finalment, s'han analitzat els resultats per a cada estil i s'han comparat amb les mesures objectives obtingudes, el que ha permès extreure algunes conclusions sobre la rellevància dels trets prosòdics en la parla expressiva, així com constatar que els resultats generats pel mòdul prosòdic han tingut una bona acceptació, encara que s'han produït diferències segons l'estil.

PARAULES CLAU: Corpus orals, prosòdia, síntesi de la parla expressiva, conversió de text a parla, tecnologies de la parla

Summary

This thesis deals with different aspects related to expressive speech synthesis (ESS). Based on the previous experience in text-to-speech (TTS) systems of the Grup en Processament Multimodal (GPMM) of Enginyeria i Arquitectura La Salle, its main aim is to improve the expressive capabilities of such systems. The expressive speech transmits paralinguistic information as, for example, the emotion of the speaker, his/her mood, a certain intention or aspects related to the environment or to his/her conversational partner. The present thesis tackles two main objectives: on the one hand, the development of an expressive speech corpus and, on the other, the modelling and the prediction of prosody from text for their use in the ESS framework.

First, an ESS system requires a speech corpus suitable for the development and the performance of some of its modules. The unavailability of a resource of this kind motivated the development of a new corpus. Based on the study of the strategies to obtain expressive speech and the previous experience of the group, the different tasks have been defined: design, recording, segmentation, tagging and validation. The main objective is to achieve a high quality speech signal and sufficient phonetic coverage (segmental and prosodic), preserving the authenticity from the point of view of the oral expressiveness. The recorded corpus has 4638 sentences and it is 5 h 12 min long; it contains five expressive styles: neutral, happy, sensual, aggressive and sad. Expressive speech has been obtained by means of the reading of texts semantically related to the defined styles. Therefore, a validation process has been required in order to guarantee that recorded utterances incorporate the desired expressive content. A comprehensive assessment of the whole corpus would be too costly. Moreover, there is insufficient scientific knowledge to completely emulate the subjective perception through automated techniques that yield a reliable validation of speech corpora. In this thesis, we propose an approach that supposes a step towards a practical solution to this problem, by combining subjective evaluation with techniques for the automatic identification of emotion in speech. The proposed method is used to perform an automatic review of the expressiveness of the corpus developed. Finally, a subjective test has allowed listeners to validate this automatic process.

Second, based on our current experience and the proposed challenges, a corpus-based system for prosody estimation has been developed. This system is characterized by modelling both the linguistic and the paralinguistic functions of prosody. A set of prosodic attributes is automatically extracted from text. This information is the input to an automatic learning system that predicts the prosodic features modelled previously by

a supervised training. The root mean squared error and the correlation coefficient have been used in both the adjustment of some system parameters and the objective evaluation. However, these measures are referred to specific utterances delivered by the speaker in the recording session, and then they do not provide information about the degree of acceptance of synthesized speech in listeners. Therefore, we have conducted different perception tests in which a group of listeners has scored a set of stimuli in each expressive style. Finally, the results for each style have been analyzed and compared with the objective measures, which has allowed to draw some conclusions about the relevance of prosodic features in expressive speech, as well as to verify that the results generated by the prosodic module have had a good acceptance, although with differences as a function of the style.

KEYWORDS: Speech corpora, prosody, expressive speech synthesis, text-to-speech, speech technology

Agradecimientos

Esta tesis doctoral no hubiese sido posible sin la ayuda de muchas personas que han sido un soporte muy fuerte a lo largo de estos últimos años. A todos ellos, mi más sincero agradecimiento y afecto.

En primer lugar quiero agradecer a mis padres la opción de vida que han hecho por sus hijos, entre los que me encuentro yo, ya que su afecto, su apoyo y el ejemplo recibido han hecho posible el camino. En especial, un recuerdo muy sentido por papá, fallecido el 8 de marzo de 2005.

A mi esposa Titina, por ser la persona que ha compartido el mayor tiempo a mi lado y porque en su compañía los momentos de debilidad se transforman en ilusión y en esperanza. Juntos hemos visto nacer y crecer a nuestros hijos que hacen posible que la soledad no exista. Además, en muchos momentos, su familia, que también es la mía, ha preferido sacrificarse para que yo pudiese disponer de ese plus de tiempo sin el cual esta tesis todavía no hubiera concluido.

A mis hermanos, porque siempre han sido para mí un ejemplo a seguir y un apoyo vital. Además quiero agradecer a mi hermana María Cinta la revisión de la ortografía y de los aspectos formales del presente trabajo.

A Joaquim Llisterri y a Joan Claudi Socoró, directores de esta tesis, por sus orientaciones y su apoyo personal, sin los cuales no hubiese sido posible la realización de este trabajo.

A Ángel Rodríguez y Patricia Lázaro por su colaboración en el desarrollo del corpus oral.

A María Jesús Machuca y Antonio Ríos por su participación en la definición de la prueba subjetiva para la evaluación de la prosodia.

A todos los compañeros del GPMM que siempre han colaborado directa o indirectamente conmigo. A todos os corresponde un pedazo de este trabajo: Francesc Alías, Rosa Mª Alsina, Germán Cobo, Lluís Formiga, David García, Xavier Gonzalvo, Elisa Martínez, Javier Melenchón, José Antonio Montero, Carlos Monzo, José Antonio Morán, Santiago Planet, Xavier Sevillano y Lluís Vicent.

A todos los alumnos y compañeros de *Enginyeria La Salle* que han colaborado

en las pruebas subjetivas de evaluación de los resultados. Un recuerdo para los alumnos que han realizado su trabajo final de carrera bajo mi supervisión, especialmente a Pere Miralles.

A las instituciones que han financiado los proyectos de I+D relacionados con esta tesis.

Finalmente, a todas aquellas personas que no he nombrado pero que también han contribuido en mi formación humana y profesional a lo largo de los años.

Índice general

Índice de figuras

Índice de tablas

Índice de algoritmos

Siglas, acrónimos y símbolos

ACR	Determinación de índices por categorías absolutas —*Absolute Category Rating*—
AFI	Alfabeto Fonético Internacional
AG	algoritmo genético
ANN	redes neuronales artificiales —*Artificial Neural Network*—
ANOVA	análisis de varianza —*ANalysis Of VAriance*—
CART	árboles de clasificación y regresión —*Classification And Regression Trees*—
CBR	razonamiento basado en casos —*Case Based Reasoning*—
CCR	Determinación de índices por categorías de comparación —*Comparison Category Rating*—
CMOS	nota media de opinión sobre las comparaciones —*Comparison Mean Opinion Score*—
CTH	conversión de texto en habla
DCR	Determinación de índices por categorías de degradación —*Degradation Category Rating*—
DMOS	nota media de opinión sobre las degradaciones —*Degradation Mean Opinion Score*—
EALS-URL	*Enginyeria i Arquitectura La Salle* de la *Universitat Ramon Llull*
F_0	Frecuencia fundamental
GA	grupo acentual
GE	grupo entonativo
GPMM	*Grup en Processament Multimodal*
HMM	Modelos ocultos de Markov

HTK	*Hidden Markov Model Toolkit*
LAICOM-UAB	Laboratorio de Análisis Instrumental de la Comunicación de la Universidad Autónoma de Barcelona
MBROLA	*MultiBand Resynthesis OverLap Add*
ML	aprendizaje automático —*Machine Learning*—
MOS	nota media de opinión —*Mean Opinion Score*—
MPEG-4	*Moving Picture Experts Group Layer-4 Video*
PLN	procesamiento del lenguaje natural
PN	prosodia natural
PS	prosodia sintética
RMSE	raíz del error cuadrático medio —*Root Mean Squared Error*—
ρ	coeficiente de correlación de Pearson
SAMPA	*Speech Assessment Methods Phonetic Alphabet*
TD-PSOLA	*Time-Domain Pitch-Synchronous Overlap and Add*
UIT	Unión Internacional de Telecomunicaciones
X-SAMPA	*eXtended Speech Assessment Methods Phonetic Alphabet*

Capítulo 1

Introducción

La presente tesis se enmarca dentro del programa de doctorado *Las TIC y su gestión* y se ha realizado en el GPMM de *Enginyeria i Arquitectura La Salle*, pertenecientes a la *Universitat Ramon Llull*, bajo la dirección de los doctores *Joan Claudi Socoró Carrié* y *Joaquim Llisterri Boix*.

1.1. Contexto

"No es lo que has dicho, sino cómo lo has dicho."

"No eres responsable de la cara que tienes, eres responsable de la cara que pones..."

Frases de la vida cotidiana de este estilo nos indican cómo las personas transmitimos actitudes, sentimientos e intenciones a través del habla y la expresión facial.

"Él era como un robot."

Sin embargo, esta frase nos indica una persona carente de afectividad, una de las cualidades humanas más esenciales. Los dos primeros ejemplos ilustran cómo las emociones están ligadas a lo que se espera en la comunicación oral humana. Este último, en cambio, sugiere que, generalmente, se considera la ausencia de emoción como una característica más propia de una máquina que de una persona. Por lo tanto, si deseamos emular el comportamiento humano con sistemas computacionales que entiendan y generen la lengua hablada, deberemos tener en cuenta el papel de la emoción en la comunicación oral y en el comportamiento humano en general.

En la actualidad, los sistemas de interacción persona-máquina (véase la figura 1.1) tienden a incorporar el habla y la visión, ya que son los canales naturales en la comunicación humana. Por esta razón, esta interacción debería ser bidireccional (Massaro et al., 2001): *i)*

la máquina podría entender el mensaje del usuario utilizando técnicas de reconocimiento automático del habla y de visión por computador (Petajan, 1984), y *ii)* la máquina podría responder mediante síntesis audiovisual (Bailly et al., 2003). Además, la interacción se volvería más eficiente y amigable si la expresión emocional pudiese reconocerse (Cowie et al., 2001) y sintetizarse (Schröder, 2001).

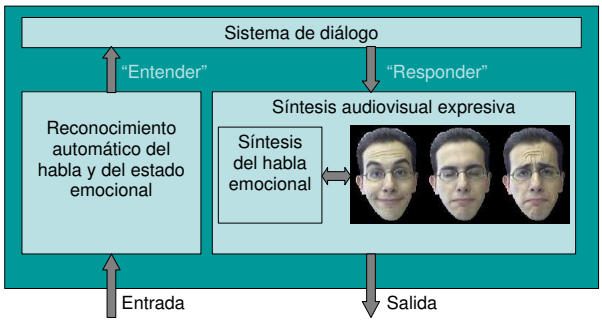


Figura 1.1: Diagrama de bloques de una interfaz persona-máquina.

Un elemento importante dentro de este contexto es la síntesis del habla expresiva, un área que presenta nuevos retos en el campo de la investigación. Desde mi punto de vista, podemos clasificar estos nuevos retos en dos categorías: los relativos a la calidad del habla generada y los relativos al desarrollo de sistemas de este tipo. De la primera categoría, destacaría que los más importantes son conseguir una mejora en la naturalidad y la expresividad. La naturalidad se define como la capacidad de generar automáticamente un habla que parezca de una persona; la expresividad la podemos definir como la capacidad de transmitir un estado de ánimo, una emoción o una intención determinada a través del habla. De la segunda categoría, la reutilización de recursos ya existentes o la reducción de costes para generar nuevas voces, nuevas emociones o nuevos estilos puede potenciar, además de la consecución del reto de mejora de la calidad, una mayor utilización de la síntesis del habla en muchas aplicaciones basadas en la interacción persona-máquina en su sentido más amplio.

1.2. Objetivos de la tesis

En este contexto, el objetivo principal de la presente tesis es el avance hacia la síntesis del habla expresiva, partiendo de la experiencia previa del GPMM en sistemas de conversión de texto en habla. La síntesis del habla expresiva comprende un área de investigación multidisciplinar que aborda uno de los problemas más complejos del procesamiento del habla y el lenguaje (Campbell et al., 2006). El habla expresiva transmite información paralingüística como por ejemplo la emoción del hablante, su estado de ánimo, una determinada intención o aspectos que le relacionan con el entorno o con su interlocutor. Los retos en este campo de investigación están relacionados con la creación de bases de datos —corpus orales—, el modelado acústico del habla expresiva (rasgos prosódicos y parámetros asociados a la cualidad de la voz), el desarrollo de sistemas de síntesis del habla y la evaluación de la calidad del habla sintetizada para una variedad de aplicaciones que no

requieran únicamente la transmisión de información lingüística.

El primer elemento necesario para poder investigar en este campo es disponer de un corpus oral adecuado para la generación de los diferentes módulos que componen un sistema de síntesis del habla de estas características. La falta de disponibilidad de un recurso de este tipo nos motivó a la producción de un nuevo corpus oral. A partir del estudio del estado de la cuestión para la consecución de habla emocionada o expresiva y la experiencia previa del grupo se plantea el diseño, la grabación, el etiquetado y la validación del nuevo corpus con el objetivo de conseguir una alta calidad de audio y una cobertura fonética suficiente, sin renunciar a la autenticidad desde el punto de vista de la expresividad oral.

La principal función del corpus que se pretende desarrollar consiste en disponer de un material para su uso en diferentes trabajos de investigación relacionados con el habla expresiva. Entre sus múltiples funciones destacan los modelados prosódicos y de la cualidad de la voz, la utilización en distintos métodos de síntesis (p.e. la síntesis basada en selección de unidades y la síntesis estadística) y la identificación automática de emociones. El diseño del corpus no se enfoca para una tarea concreta de síntesis del habla sino que se prioriza la obtención de una determinada diversidad expresiva y de un volumen suficiente de datos.

El segundo objetivo ha consistido en el desarrollo de un método para predecir la prosodia a partir del texto dentro del contexto de la síntesis del habla expresiva. Se pretende aprender de forma automática y conjunta las funciones lingüística y paralingüística (aquella que complementa el mensaje con una intención determinada o que refleja una actitud o estado emocional del hablante) para diferentes estilos expresivos. El corpus obtenido mediante la consecución del primer objetivo será utilizado para la investigación planteada en este segundo objetivo.

1.3. Contenidos de la tesis

La tesis comienza con el capítulo 2 en el cual se presentan una serie de fundamentos que abarcan diferentes disciplinas y que tienen relación con el ámbito de esta tesis, tales como la teoría de las emociones, su expresión y percepción y, por último, unas nociones sobre la conversión de texto en habla.

A continuación se expone, en el capítulo 3, el estado de la cuestión de los dos elementos clave para la presente investigación: *i)* los corpus orales para el estudio y el desarrollo de aplicaciones relacionadas con el habla expresiva; y *ii)* los dos elementos que intervienen principalmente en la síntesis del habla expresiva que son el modelado prosódico y los métodos de síntesis que se pueden aplicar.

El capítulo 4 trata el proceso completo de la producción de un corpus del habla orientado a la síntesis expresiva del habla, desde su diseño hasta su validación final. Se dedica un especial énfasis a la validación del corpus desde el punto de vista de la expresividad utilizando técnicas de identificación automática de la emoción a partir del habla. La

principal aportación en el ámbito de los corpus de habla expresiva es la propuesta de un método de revisión automática de todas las locuciones del corpus para verificar su contenido emocional. Este método está guiado por los resultados de una primera prueba subjetiva de identificación de emociones realizada con una muestra del corpus, que además, ha permitido constatar, de forma general, que el contenido expresivo del corpus es bueno. De todas formas, al tratarse de un corpus grabado por una locutora, se requiere una revisión completa en la que se detecte qué locuciones carecen de la expresividad deseada. Dado que el tamaño del corpus no permite una revisión manual exhaustiva, se ha propuesto un método automático que una vez aplicado a todo el corpus, ha sido validado mediante una segunda prueba de percepción con oyentes.

En el capítulo 5 se presentan las diferentes fases que se han seguido para desarrollar y evaluar un módulo de generación automática de la prosodia en el ámbito de la síntesis del habla expresiva. El capítulo comienza con la descripción de dos investigaciones preliminares que han servido de base para el desarrollo del sistema final. El sistema de modelado y estimación de la prosodia se basa en el razonamiento basado en casos —una técnica de aprendizaje automático por analogía— que se ha evaluado utilizando el corpus de habla expresiva descrito en el capítulo 4. En la fase de ajuste del sistema (entrenamiento) se han utilizado medidas objetivas del error y la correlación respecto un conjunto de locuciones dedicadas a la fase de test. Además, se ha llevado a cabo una prueba de escucha con oyentes que han puntuado una serie de estímulos de cada estilo. Los resultados de la prueba subjetiva se analizan para cada estilo y se comparan con las medidas objetivas obtenidas. Estos resultados permiten tener una medida del grado de aceptación del habla sintetizada respecto a ejemplos de habla natural.

Finalmente, en el capítulo 6 se exponen las principales conclusiones, así como las líneas futuras de trabajo que se abren y que dan continuidad a las aportaciones de este trabajo de tesis.

Capítulo 2

Fundamentos

En este capítulo se abordan materias de diferentes disciplinas relacionadas con el desarrollo de la presente tesis. Básicamente, se revisa el concepto de emoción y su representación (apartado 2.1), ya que es uno de los elementos más íntimamente relacionados con la expresión humana. A continuación, se describe la base de la expresión y la percepción de las emociones (apartado 2.2). Por último, en un plano más tecnológico, se realiza una introducción a la conversión de texto en habla (apartado 2.3).

2.1. Teoría de las emociones

2.1.1. El concepto de emoción

La palabra "emoción" tiene diferentes significados según el ámbito en el que se emplee. La metodología para describir la emoción presenta muchas variantes en función de la disciplina, siguiendo un amplio recorrido desde la biología hasta la psicología. En este apartado veremos las diferentes acepciones de esta palabra en el ámbito de la investigación que ocupa este trabajo, "habla y emoción" (del inglés *speech and emotion*), y que busca las relaciones entre estos dos dominios.

Cowie y Cornelius (2003) presentan un estudio exhaustivo de términos y conceptos relacionados con la emoción y el habla. En él, se anima a la comunidad de investigadores en tecnologías del habla a abordar el tema de la descripción de las emociones sin esperar una solución completa aportada desde otra disciplina. Un primer término que tratan es el de "emoción plena" (Scherer, 1999), al que otros autores se refieren como "emociones primarias" (Plutchik, 2001) o "emociones básicas" (Ekman, 1999). Con estos términos se denota la forma más intensa de las emociones. En este caso, están presentes todos los aspectos considerados relevantes de una emoción en concreto, tales como la evaluación de la situación, los acontecimientos previos, la respuesta conductual, los aspectos psicológicos y las señales universales distintivas.

En segundo lugar, se presenta el término de "emoción subyacente", que denota una clase de colorido emocional presente en todos los estados mentales. La descripción de estas emociones subyacentes no es fácil, pero es cierto que en la comunicación humana aparecen mucho más a menudo que las emociones plenas.

Finalmente, se decide denominar "estados emocionales" a toda la variedad de estados que van desde las emociones subyacentes más débiles hasta las emociones plenas. Este abanico contiene todo un conjunto de estados intermedios que tienen sentido en el ámbito de la comunicación humana.

Además se introduce el concepto de estados relacionados con la emoción en los cuales las personas no sienten propiamente una emoción, pero presentan ciertos aspectos propios de las emociones (humor, excitación, cierta actitud, etc.).

2.1.2. Teorías sobre emociones plenas

Scherer (1986) describió la emoción como "la interfaz del organismo hacia el mundo exterior", destacando tres funciones principales de la emociones:

- Reflejan la evaluación de la relevancia y el significado del estímulo particular en términos de las necesidades del organismo, planes y preferencias (valoración de la situación).
- Preparan fisiológica y psicológicamente al organismo para una acción apropiada (cambios fisiológicos y tendencia a la acción).
- Comunican el estado del organismo y las intenciones de comportamiento hacia otros seres próximos (comportamiento expresivo facial, corporal y oral).

Las teorías contemporáneas sobre la emoción en la psicología, revelan cuatro perspectivas básicas, que comprenden desde las primeras aproximaciones de Charles Darwin, hasta las teorías de finales del siglo XX, sobre cómo definir, estudiar y explicar las emociones.

Cornelius (2000) define estas cuatro perspectivas como: Darwiniana, Jamesiana, Cognitiva y Constructivista Social. Cada una de ellas se basa en sus propias suposiciones sobre cómo construir teorías sobre la emoción, la naturaleza de la misma, y sobre cómo dirigir la investigación de las emociones. Aún así, hay coincidencias destacables entre las cuatro teorías, sobre todo entre la Darwiniana y la Jamesiana.

2.1.2.1. La Perspectiva Darwiniana

La idea básica de la perspectiva Darwiniana es que las emociones son fenómenos desarrollados como funciones importantes de supervivencia, seleccionadas como tal para solucionar ciertos problemas a los que la especie humana ha tenido que hacer frente. Por

ello, los comportamientos emocionales son similares en todos los seres humanos e incluso a los de aquellos mamíferos con los que el hombre ha compartido un pasado a lo largo de la evolución. Los inicios de esta perspectiva se remontan al año 1872, con el libro de Charles Darwin *The Expression of Emotion in Man and Animals*[1]. Sus ideas han sido muy influyentes. Su legado en el estudio de la emoción en la psicología y la biología se basa en:

- Aplicar sus teorías de la evolución por selección natural con el fin de entender las expresiones emocionales y, por extensión, las propias emociones.
- Remarcar que las expresiones emocionales tienen que entenderse en términos de sus funciones y, por lo tanto, como un valor de supervivencia.

2.1.2.2. La Perspectiva Jamesiana

La perspectiva Jamesiana, fue inspirada por los escritos de William James sobre la emoción (*What is an emotion?, 1884)*[2], de los cuales destaca su famosa ecuación sobre las relaciones entre las emociones y los cambios corporales: "Los cambios corporales siguen directamente la percepción de una excitación y, la emoción es el sentimiento experimentado al aparecer estos mismos cambios."(traducido de James, 1884, págs. 189-190). James insiste en que sería imposible tener emociones sin que aparecieran cambios corporales, y en que estos cambios siempre aparecen antes que la emoción. James tomaba como eje central de sus estudios, la explicación de la naturaleza propia de las emociones, mientras que Darwin se centraba en sus manifestaciones. Aun así, ambos coincidían en que las emociones eran adaptaciones al entorno y que tenían importantes funciones relacionadas con la supervivencia. Según esta perspectiva, el hombre experimenta emociones debido a que el cuerpo ha aprendido a responder, automática y evolutivamente, a las características del entorno. El cuerpo responde primero, y nuestra experiencia a los cambios constituye lo que se denomina emoción. James escribió: "Estamos tristes porque lloramos, enfadados porque golpeamos y tenemos miedo porque temblamos."(traducido de James, 1884, pág. 190). No queda claramente definido cómo los cambios corporales son iniciados por la percepción de los acontecimientos ambientales, y tal cuestión no se resolverá hasta la denominada revolución cognitiva de la psicología.

2.1.2.3. La Perspectiva Cognitiva

La perspectiva cognitiva es la más dominante de las cuatro, y esto es así gracias a que esta perspectiva ha sido minuciosamente incorporada dentro de las otras tres. La aproximación cognitiva moderna se basa en los estudios de las emociones realizados por Magda Arnold, pero los orígenes de la misma datan más allá de los filósofos helenísticos. El eje central de esta perspectiva es que la emoción y el pensamiento son inseparables;

[1] La web `http://darwin-online.org.uk/` contiene la obra principal de Charles Darwin.

[2] La web `http://psychclassics.yorku.ca/` permite la consulta en línea de esta obra, así como una amplia extensión de obras clásicas de la psicología.

más específicamente, todas las emociones son enjuiciadas mediante una evaluación. Este proceso de evaluación consiste en discernir qué acontecimientos del entorno son tomados como buenos o malos por nosotros. Arnold criticó a James por no especificar cómo aparecían los cambios corporales ante la percepción de los acontecimientos ambientales. Para Arnold, la conexión perdida es el proceso de evaluación. Del mismo modo que James no podía concebir una emoción sin un cuerpo, Arnold, no lo podía hacer sin una evaluación. Cada emoción está asociada a un patrón específico y diferente de evaluación. Estos patrones proporcionan la conexión entre las características particulares de la persona, su aprendizaje, el temperamento, la personalidad, el estado psicológico y las características particulares de la situación en que se encuentra la persona. El proceso de evaluación informa al organismo de las características particulares del entorno y proporciona la manera de actuar frente a estas.

2.1.2.4. La Perspectiva Constructivista Social

De las cuatro perspectivas, esta es la más joven, diversa y la que genera más controversia. Rompiendo los esquemas de quienes ven las emociones como una adaptación al medio, los constructivistas sociales creen que las emociones son productos culturales fijados por las reglas sociales adoptadas. Según James Averill "las emociones no son remanentes de nuestro pasado psico-genético, ni pueden ser explicadas en términos estrictamente psicológicos. Más bien, son construcciones sociales, y estas, solo pueden ser plenamente entendidas a partir de un análisis social" (traducción de Averill, 1980, citado por Cornelius, 2000, pág. 5). Para los constructivistas sociales la cultura juega un papel central en la organización de las emociones, ya que es la que determina el proceso de evaluación mediante reglas sociales.

2.1.3. La descripción de las emociones

2.1.3.1. Emociones básicas

Muchas de las teorías sobre la emoción, especialmente las que siguen las tradiciones Darwiniana y Jamesiana, utilizan el concepto de emociones básicas, a partir de las cuales se generan todas las demás mediante variaciones o combinaciones de estas. No hay un criterio único para definir qué emociones forman este conjunto básico. Las 4 emociones básicas más aceptadas son: alegría, tristeza, enfado y miedo, que se considera que están directamente ligadas a procesos biológicos. La mayor parte de teorías coinciden en que hay un número inferior a diez emociones básicas, aunque estudios más recientes (Cowie y Cornelius, 2003) definen entre 10 y 20 (véase la tabla 2.1). Cabe destacar el término "The Big Six" utilizado en Cornelius (2000), en el que se engloba al conjunto formado por la felicidad (*happiness*), la tristeza (*sadness*), el miedo (*fear*), el asco (*disgust*), el enfado (*anger*) y la sorpresa (*surprise*).

Las emociones que forman parte de estos conjuntos, se denominan emociones plenas, básicas o primarias, y se consideran fundamentales, puesto que representan los patrones

relacionados con la supervivencia, es decir, las respuestas a acontecimientos que han sido seleccionadas a lo largo de la historia de la evolución y, además, porque todo el resto de emociones derivan de estas.

Hay que tener en cuenta que uno de los problemas en la investigación intercultural es la traducción precisa de los términos relacionados con la emoción. Debido a la connotación de cada término, no hay una solución satisfactoria a este problema. Scherer (1988) presenta una lista de descriptores de la emoción en 5 lenguas indoeuropeas, fruto de la actividad de investigación de un equipo de psicólogos de diferentes países. La traducción de los términos de la tabla 2.1 se ha hecho con la ayuda de este estudio. La columna traducida de la derecha, que puede representar la unión de los conjuntos de emociones básicas de los 6 estudios analizados, nos proporciona un nuevo conjunto de 34 emociones básicas, con lo que se puede concluir que la representación del espacio emocional mediante emociones discretas es demasiado compleja para su utilización en aplicaciones prácticas.

Ekman (1999) propone el concepto de familias de emociones, ya que considera que cada emoción no es un único estado afectivo, sino una familia de estados relacionados. Cada familia se caracteriza por un tema, fruto de la evolución, y unas variaciones, reflejo del aprendizaje. Se propone una lista con 15 emociones básicas (o familias) mostradas en la segunda columna de la tabla 2.1. Por ejemplo la familia *Anger* abarcaría emociones como enojo, enfado y rabia, todas con un tema común, pero con diferentes matices fruto de elementos adquiridos previamente.

Cabe destacar que existe un número importante de términos que describen 'estados relacionados con la emoción' (por ejemplo, confiado, relajado, aburrido, etc.), cosa que refleja el sentido generalizado de que estos constituyen una parte significativa de la vida emocional diaria. Por tanto, podemos concluir que las teorías sobre las emociones mayoritariamente se refieren a emociones básicas y consideran que las demás emociones son combinaciones o modificaciones de estas emociones básicas, aunque no hay un consenso claro sobre qué emociones son las llamadas emociones básicas.

2.1.3.2. Modelos circunflejos

Algunos investigadores han concluido que las emociones se pueden representar mediante una estructura circular. La proximidad de dos categorías representa emociones conceptualmente similares, mientras que las emociones contrarias están separadas 180 grados. El primer modelo circunflejo es obra de Harold Schlosberg (1941) obtenido al observar que los errores de reconocimiento de la expresión facial se correspondían a la confusión entre categorías adyacentes situadas sobre una circunferencia (Schröder, 2004, p. 25). En 1958, Robert Plutchik propuso un modelo con 8 emociones básicas bipolares: alegría-aburrimiento, enfado-miedo, aceptación-asco, y sorpresa-expectación. Una evolución de esta teoría ha llevado al modelo circunflejo tridimensional (Plutchik, 2001) en el cual se representan 4 aspectos:

- La dimensión vertical representa la intensidad.

Tabla 2.1: Listas recientes de emociones básicas reproducidas de Cowie y Cornelius (2003) y su traducción al español

Lazarus (1999)	Ekman (1999)	Buck (1999)	Lewis-Haviland (1993)	Banse-Scherer (1996)	Cowie et al. (1999)	Traducción
Anger	Anger	Anger	Anger/hostility	Rage/hot anger Irritation/cold anger	Angry	Enfado/enfadado
Fright	Fear	Fear	Fear	Fear/terror	Afraid	Miedo/atemorizado
Sadness	Sadness/distress	Sadness	Sadness	Sadness/dejection Grief/desperation	Sad	Tristeza/triste
Anxiety		Anxiety	Anxiety	Worry/anxiety	Worried	Inquietud/preocupado
Happiness	Sensory pleasure	Happiness	Happiness	Happiness Elation (joy)	Happy	Alegría/Felicidad/feliz
	Amusement		Humour		Amused	Diversión/divertido
	Satisfaction				Pleased	Satisfacción/satisfecho
	Contentment				Content	Contento
		Interested			Interested	Interesado
		Curious				Curioso
		Surprised				Sorprendido
	Excitement				Excited	Ilusión/excitado
		Bored		Boredom/indifference	Bored	Aburrido
					Relaxed	Relajado
		Burnt out				Quemado/estresado
Disgust	Disgust	Disgust	Disgust	Disgust		Asco
	Contempt	Scorn		Contempt/scorn		Desprecio/desdén
Pride	Pride	Pride	Pride			Orgullo
		Arrogance				Arrogancia
Jealousy		Jealousy				Celos
Envy		Envy				Envidia
Shame	Shame	Shame	Shame	Shame/guilty		Vergüenza
Guilt	Guilt	Guilt	Guilt			Culpabilidad
	Embarrassment		Embarrassment			Desconcierto
					Disappointed	Desilusionado
Relief	Relief					Alivio
Hope						Esperanza
					Confident	Confiado
Love			Love		Loving	Amor/cariñoso
					Affectionate	Afectuoso
Compassion		Pity				Compasión
		Moral rapture				Entusiasmo
		Moral indignation				Indignación
Aesthetic						Estético

- El círculo representa grados de similitud entre emociones, de forma que las emociones similares están próximas y las opuestas están separadas 180 grados.
- Los ocho sectores representan las emociones básicas bipolares.
- Los espacios en blanco representan emociones que son mezclas de dos emociones primarias, por ejemplo el asco y la rabia producen odio.

Si analizamos la figura 2.1, podemos observar que las emociones secundarias se producen por combinación de emociones primarias adyacentes. Por ejemplo, el remordimiento se concibe como una mezcla de tristeza y aversión hacia la propia conducta. Además, variando la intensidad de las emociones se pueden obtener nuevos estados emocionales. Por ejemplo, el temor puede ir desde una simple aprehensión hasta un enorme terror. También muestra que hay emociones opuestas que, por lo tanto, no se pueden mezclar, como la tristeza y la felicidad.

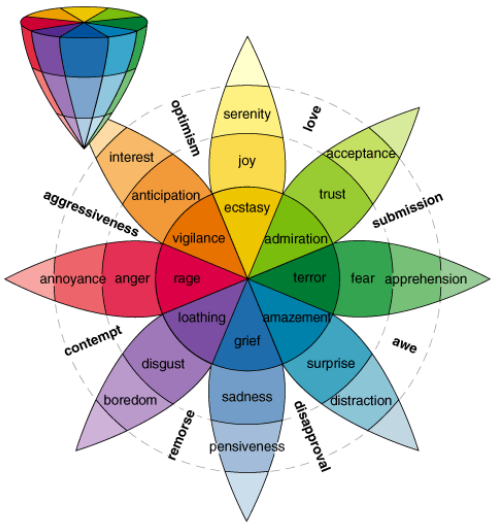


Figura 2.1: Modelo circunflejo tridimensional de Plutchik (2001)

2.1.3.3. Espacio multidimensional

Un objetivo fundamental para una descripción sistemática de las emociones consiste en encontrar formas de representar los estados emocionales como coordenadas en un espacio con un número pequeño de dimensiones. Se han llevado a cabo numerosas investigaciones que abordan este tipo de clasificación. Una revisión histórica de diferentes definiciones de estas dimensiones se puede encontrar en Cowie y Cornelius (2003). Un estudio más amplio sobre la descripción de los estados emocionales mediante espacios multidimensionales se presenta en Schröder (2004), comenzando por una perspectiva histórica y continuando con una descripción del significado de estas dimensiones y qué relación tienen con el comportamiento humano.

De forma resumida, la mayor parte de los estudios intentan representar el espacio emocional en dos dimensiones, aunque algunos añaden una tercera. La terminología asociada a cada eje también presenta diferencias según el estudio. A continuación se presentan las tres dimensiones más utilizadas junto con diferentes términos para referirse a ellas:

- Evaluación / agrado / valoración: corresponde al eje "Positivo-Negativo", que clasifica las emociones según lo placentero o desagradable de estas (p. ej. desde la alegría hasta el enfado).
- Activación / actividad: corresponde a una escala "Activo-Pasivo", que indica la presencia o ausencia de energía o tensión. (p. ej., desde estar furioso a estar aburrido).
- Potencia / fuerza: corresponde a la escala "Dominante-Sumiso", distinguiendo emociones iniciadas por el sujeto de aquellas causadas por el entorno (p. ej., desde el desprecio al temor o la sorpresa).

Las emociones con una actividad similar, como la alegría y el enfado, se confunden más entre sí, que emociones con valoración o fuerza semejante.

Esta representación del espacio emocional es muy utilizada, destacando la herramienta *Feeltrace* (véase la figura 2.2), que permite el etiquetado en dos dimensiones emocionales: evaluación (eje horizontal) y activación (eje vertical). Esta herramienta permite la anotación del estado emocional percibido marcando puntos a medida que avanza la reproducción del audio o el vídeo seleccionado. En Cowie et al. (2000a), se destaca que esta herramienta tiene la misma potencia que un vocabulario emocional con 20 palabras no superpuestas, pero además tiene la ventaja de permitir estados intermedios y de representar la evolución temporal de un estado emocional a otro.

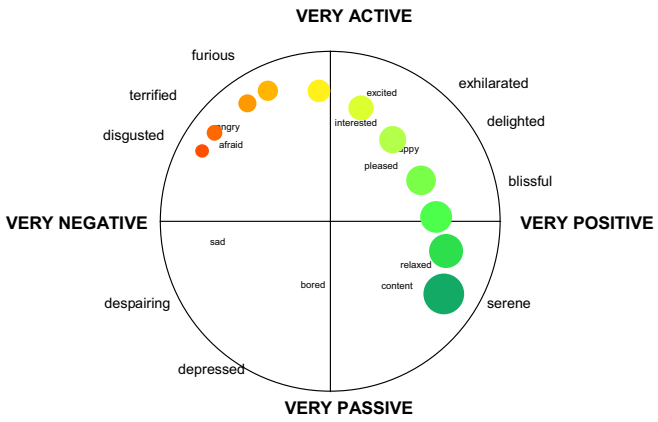


Figura 2.2: Imagen de la pantalla de la herramienta *Feeltrace* (Cowie et al., 2000a) utilizada para anotar la emoción de un estímulo sonoro o visual en una escala bidimensional.

2.2. Expresión y percepción de emociones

La expresión de emociones es un hecho habitual para las personas, ya que es uno de los elementos más importantes de la comunicación humana. Numerosos estudios han tratado de averiguar los efectos de la emoción en diferentes ámbitos del comportamiento humano (puede verse un resumen en Bartneck, 2000). Los efectos que se pueden percibir serán muy importantes desde el punto de vista comunicativo. Podemos hablar de efectos fisiológicos, en el habla (acústicos, prosódicos y léxicos) y en el lenguaje corporal (gestos, expresión facial y movimientos corporales). Se parte de la hipótesis que la voz sufre cambios acústicos causados directamente por alteraciones fisiológicas cuando una persona se encuentra en un determinado estado emocional (Scherer, 1986). Por ejemplo, una activación del sistema nervioso simpático ocurre cuando sentimos rabia o miedo, provocando cambios en el organismo como un incremento de la presión arterial o de la frecuencia cardíaca, temblores, sequedad de boca, etc. Estos cambios fisiológicos provocan cambios en el habla y en la expresión facial. Por lo tanto, la investigación en el campo de la expresividad emocional, requerirá de modelos acústicos consistentes en la definición de los parámetros del habla y su cuantificación para cada estado emocional. En los dos apartados siguientes se hace una breve introducción a la expresividad vocal y a la interpretación musical como fuentes para la comunicación de estados emocionales.

2.2.1. Parámetros del habla relacionados con la emoción

El habla por sí sola es un elemento suficiente para comunicar emociones. Por ejemplo en una conversación telefónica podemos captar el enfado del interlocutor sólo por el tono de su voz. De hecho, los oyentes esperan un cierto grado de emoción en la voz como parte esencial del habla humana. El componente expresivo o afectivo del habla es principalmente no-léxico, aunque hay que tener en cuenta otros elementos importantes de la comunicación: el contexto, el contenido del mensaje, los gestos y la expresión facial —si se da el caso.

La principal fuente de energía en el habla es la vibración de los pliegues vocales. Para un instante dado, la velocidad de vibración de los pliegues vocales determina la frecuencia fundamental de la señal acústica. Esta vibración de los pliegues vocales genera un espectro de armónicos (frecuencias múltiples de la fundamental) que, al ser filtrados por el tracto vocal, producen los diferentes sonidos. Hay que tener en cuenta que también existen fuentes de energía aperiódicas —continuas o impulsionales— provocadas por la fricción del paso del aire por diferentes zonas del tracto vocal o debidas a un cierre seguido de una explosión. La evolución temporal de estas fuentes de energía y del tracto vocal generan una onda acústica que representamos mediante la señal de voz.

Las variaciones en la intensidad y la frecuencia fundamental, la duración de los sonidos del habla, la posición y la duración de las pausas son los principales rasgos prosódicos del habla (Llisterri et al., 2004). La prosodia tiene principalmente una función lingüística, como por ejemplo distinguir entre una afirmación o una pregunta. Además, el habla presenta variaciones en sus rasgos prosódicos y en el timbre que no son relevantes desde el punto de vista estrictamente lingüístico. En este caso, distinguimos entre la función

paralingüística, que complementa el mensaje con una intención determinada o que refleja una actitud o estado emocional del hablante, y la función extralingüística que aporta información sobre las características del locutor, como su edad, su sexo, su estatus socioeconómico, etc. (Escudero, 2003).

A continuación se definen las propiedades acústicas de los sonidos del habla relacionadas con la expresividad vocal a las que se hará referencia posteriormente:

1. Propiedades relacionadas con la **melodía**[3]:
   - **Frecuencia fundamental ($\mathbf{F_0}$)**: Resultado de la vibración de los pliegues vocales que se define como el ciclo periódico de la señal de voz. Su medida habitual es el hercio (Hz) que mide los ciclos por segundo.
   - **Curva de $\mathbf{F_0}$ o melódica**: Es la secuencia de valores de F_0 para una elocución y está relacionada con la percepción de la entonación del habla.
   - ***Jitter***: Perturbación a pequeña escala en la F_0, debida a fluctuaciones en los tiempos de apertura y cierre de los pliegues vocales de un ciclo al siguiente.
2. Propiedades relacionadas con la **intensidad**[4]:
   - **Intensidad**: Medida de la energía de la señal acústica. Habitualmente se utiliza una transformación logarítmica de la amplitud de la señal llamada decibelio (dB) que representa mejor la percepción humana del sonido.
   - ***Shimmer***: Perturbación a pequeña escala en la intensidad debida a fluctuaciones en la amplitud de un ciclo al siguiente.
3. Propiedades relacionadas con los **aspectos temporales** del habla:
   - **Velocidad del habla**: Se mide a partir de la duración de los segmentos del habla o como el número de unidades lingüísticas por unidad temporal (e.g. palabras por minuto o sílabas por segundo).
   - **Pausas**: Habitualmente se mide el número y la duración de los silencios en la señal de voz[5].
4. Propiedades relacionadas con el **timbre**:
   - **Energía de alta frecuencia**: Proporción relativa de la energía por encima de una frecuencia de corte respecto a la energía total.

[3]Según Garrido (1991), la melodía (*pitch* en inglés) es el fenómeno que se relaciona con la curva de Frecuencia fundamental (F_0) o curva melódica de un grupo fónico. No hay que confundirla con la entonación que es un fenómeno lingüístico relacionado con la sensación perceptiva que produce la variación de tres parámetros físicos: F_0, amplitud y duración.

[4]El principal correlato perceptivo de la intensidad es la sonía (*loudness* en inglés) que está relacionada con el nivel de sensación sonora.

[5]En este caso, se trata de pausas vacías (*empty pauses* en inglés) que se realizan para respirar. También existen pausas llenas (*filled pauses* en inglés) en las que sí existe producción sonora y se relacionan con la planificación del discurso (Puigví et al., 1994).

- **Frecuencias de los formantes**: Se trata de regiones de frecuencia que presentan una alta concentración de energía espectral, y que reflejan las resonancias naturales del tracto vocal. Se suelen representar por la frecuencia central de la región y su ancho de banda.
- **Precisión en la articulación**: Mide la desviación de las frecuencias de los formantes en las vocales desde las frecuencias formantes neutras (Juslin y Laukka, 2003).

Habitualmente el *jitter* y el *shimmer* no se asocian a propiedades de la prosodia aunque están relacionadas con la F_0 y la intensidad respectivamente, sino que se suelen agrupar junto con las propiedades del timbre. A este conjunto de propiedades que forman las perturbaciones de la F_0 y de la intensidad más las propiedades del timbre nos referiremos como **cualidad de la voz** (del inglés, *voice quality*).

Se han publicado numerosos estudios sobre la correlación entre habla y emoción. Murray y Arnott (1993) presentaron un resumen de los trabajos más significativos en la bibliografía sobre emoción y habla. Concluyeron que la mayor parte de los estudios coincidían en los efectos vocales de algunas emociones. De hecho, distinguieron entre emociones primarias (rabia, alegría, tristeza, miedo y asco) y emociones secundarias (pena, ternura, ironía, sorpresa). La tabla 2.2 muestra una traducción del conocido resumen de Murray y Arnott (1993), en el que se describen los efectos más comúnmente asociados a las emociones indicadas y que están descritos respecto a un estilo de habla neutro.

Tabla 2.2: Resumen de los efectos de las emociones en el habla, traducido de Murray y Arnott (1993)

	Miedo	Alegría	Tristeza	Enfado	Asco
Velocidad del habla	Ligeramente más rápida	Más rápida o más lenta	Ligeramente más lenta	Mucho más rápida	Mucho más lenta
Promedio de F_0	Mucho más alta	Más alta	Ligeramente más baja	Mucho más alta	Mucho más baja
Rango de F_0	Más amplio	Más amplio	Ligeramente más estrecho	Más amplio	Ligeramente más amplio
Cualidad de la voz	Jadeante	Estrepitosa	Resonante	Sonoridad irregular	Ruidosa
Cambios de F_0	Abruptos en sílabas tónicas	Suaves inflexiones ascendentes	Inflexiones descendentes	Normal	Amplios en inflexiones descendentes finales
Articulación	Tensa	Normal	Arrastrada	Precisa	Normal

La cuantificación de los parámetros del habla en esta tabla es imprecisa. Para la obtención de modelos acústicos de las emociones se necesitan enfoques con un mayor nivel de precisión en la cuantificación. En Cowie et al. (2001) se presenta un amplio estudio sobre habla y emoción. Más concretamente, se incluye un resumen que cubre la mayor parte del material disponible hasta la fecha sobre las características del habla con emociones específicas. La tabla está formada por 14 estados emocionales caracterizados

por una descripción cualitativa de las características del habla organizadas en 5 categorías (acústica, contorno melódico, tono, cualidad de la voz y otros).

Otro trabajo muy completo es el presentado por Juslin y Laukka (2003), en el que analizan 104 estudios relacionados con la expresión vocal y 41 estudios sobre interpretación musical, con el objetivo de descubrir si las dos modalidades comunican las emociones de forma similar. El estudio se ha centrado en cinco categorías emocionales: enfado, miedo, alegría, tristeza y amor-ternura. A partir del análisis comparativo de diferentes trabajos, se ha estudiado si la identificación emocional a través del habla es un fenómeno transcultural. Se concluye que la tristeza y el enfado son las emociones mejor decodificadas, tanto en los estudios intraculturales como en los transculturales. En cambio, la alegría es menos identificada entre culturas distintas, a diferencia de lo que sucede en los estudios sobre la expresión facial.

Otra contribución importante de Juslin y Laukka (2003) es la recopilación de los indicadores acústicos más utilizados para expresar emociones representadas de forma discreta. La dificultad de comparar estudios con datos cuantitativos no uniformes se ha solucionado agrupando los resultados en categorías más amplias (p.ej. alto, medio, bajo). La tabla 2.3 muestra un resumen de los parámetros acústicos más analizados según esta recopilación acerca de estudios sobre la expresión vocal. Se muestra la categoría mayoritaria para cada par parámetro-emoción, indicando entre paréntesis el número de estudios que se catalogan según la categoría indicada respecto a todos los estudios que han tratado dicho par. Se puede comprobar que el parámetro mayoritariamente analizado es el valor medio de la F_0, seguido de su variabilidad. Si se comparan estos resultados con los de la tabla 2.2 para las cuatro emociones coincidentes (enfado, miedo, alegría y tristeza), se observa que hay plena coincidencia excepto para el par "Variabilidad de F_0-Miedo", en el que los resultados son contrarios. De todas formas, Juslin y Laukka (2003) ya detectan cierta contradicción para este caso, porque de los 32 estudios analizados, nueve de ellos consideran una alta variabilidad de F_0 y otros seis proponen una variabilidad media. El siguiente parámetro más estudiado es la velocidad del habla, para el cual hay plena concordancia de los dos estudios en las cuatro emociones. El tercer conjunto de parámetros más estudiado es el que hace referencia a la intensidad. Por lo tanto, los rasgos prosódicos F_0, velocidad del habla e intensidad (en este orden) son los más analizados en los estudios de expresión vocal. Se detecta una menor presencia de estudios que consideran parámetros relacionados con la cualidad de la voz.

Tabla 2.3: Resumen de los indicadores vocales utilizados para expresar emociones discretas extraídos de diferentes estudios de expresión vocal según Juslin y Laukka (2003). Se muestran los parámetros del habla más representativos, indicando entre paréntesis el número de estudios que obtienen la categoría para el par parámetro-emoción correspondiente respecto al total de los estudios que han estudiado este par concreto.

	Enfado	Miedo	Alegría	Tristeza	Ternura
Media de F0	Alta (33/43)	Alta (28/39)	Alta (34/38)	Baja (40/45)	Baja (4/5)
Variabilidad de F0	Alta (27/35)	Baja (17/32)	Alta (33/36)	Baja (31/34)	Baja (5/5)
Contorno de F0	Ascendente (6/8)	Ascendente (6/6)	Ascendente (7/7)	Descendente (11/11)	- (0/0)
Perturbación de F0 (*Jitter*)	Alta (6/7)	*Empate* (4/8)	Alta (5/8)	Baja (5/6)	- (0/0)
Media de Intensidad	Alta (30/32)	Alta (11/22)	Alta (20/26)	Baja (29/32)	Baja (4/4)
Variabilidad de Intensidad	Alta (30/32)	Alta (11/22)	Alta (20/26)	Baja (29/32)	Baja (4/4)
Energía de alta frecuencia	Alta (22/22)	Alta (8/16)	Alta (13/17)	Baja (19/19)	Baja (3/3)
Velocidad del habla	Rápida (28/35)	Rápida (24/29)	Rápida (22/33)	Lenta (30/36)	Lenta (3/4)
Proporción de pausas	Pequeña (8/8)	Pequeña (4/9)	Pequeña (3/6)	Grande (11/12)	Grande (1/1)
Precisión en la articulación	Alta (7/7)	*Empate* (2/6)	Alta (3/5)	Baja (6/6)	Baja (1/1)
Media de la frec. del 1er Formante	Alta (6/6)	Baja (3/4)	Alta (5/6)	Baja (5/6)	- (0/0)

2.2.2. La interpretación musical

En Bartneck (2000) se presenta un resumen de los estudios más relevantes referidos a la expresión emocional a través de la música. El modelado acústico de dicha expresión no es una tarea fácil ya que depende de hechos culturales, las habilidades del intérprete, la diferente percepción en función del oyente, etc. Los atributos emocionales de la música están mayoritariamente presentes en la manipulación de la amplitud, el tono (nivel, variación y contorno), el *tempo* y el timbre. Hay una cierta superposición con los resultados obtenidos para el habla emocional, como se deduce del trabajo comparativo de Juslin y Laukka (2003). A modo de ejemplo, reproducimos una tabla que resume esta similitud entre la expresividad vocal y la interpretación musical (véase la tabla 2.4).

Tabla 2.4: Resumen de las propiedades acústicas que presentan un patrón de comportamiento parecido para la expresión vocal y la interpretación musical en cuatro emociones según Juslin y Laukka (2003).

Emoción	Propiedades acústicas (expresión vocal/interpretación musical)
Enfado	Velocidad/*tempo* rápida/o, intensidad/nivel de sonido fuerte, alta variabilidad intensidad/nivel de sonido, alta energía de alta frecuencia, alto nivel F0/tono, alta variabilidad F0/tono, contorno ascendente F0/tono, rápido inicio de voz/ataque
Miedo	Velocidad/*tempo* rápida/o, intensidad/nivel de sonido baja/o (excepto en pánico), alta variabilidad intensidad/nivel de sonido, baja energía de alta frecuencia, alto nivel F0/tono, poca variabilidad F0/tono, contorno ascendente F0/tono
Alegría	Velocidad/*tempo* rápida/o, intensidad/nivel de sonido media/o-fuerte, energía media de alta frecuencia, alto nivel F0/tono, alta variabilidad F0/tono, contorno ascendente F0/tono, rápido inicio de voz/ataque
Tristeza	Velocidad/*tempo* lenta/o, intensidad/nivel de sonido baja/o, poca variabilidad intensidad/nivel de sonido, pequeña energía de alta frecuencia, bajo nivel F0/tono, poca variabilidad F0/tono, contorno descendente F0/tono, lento inicio de voz/ataque

2.3. Conversión de texto en habla

La conversión de texto en habla (CTH) consiste en la transformación de un texto cualquiera en su equivalente sonoro. Durante este proceso de transformación, el texto de entrada pasa por una serie de módulos que van añadiendo nueva información necesaria para la correcta lectura del texto. El primer requisito de un sistema de CTH es conseguir una elevada inteligibilidad, es decir, las palabras generadas deben ser claramente identificables por los oyentes. En la década de los noventa, los sistemas de síntesis concatenativa basada en *Time-Domain Pitch-Synchronous Overlap and Add* (TD-PSOLA), desarrollados a partir de las técnicas descritas por Moulines y Charpentier (1990), consiguieron altas tasas de inteligibilidad (Dutoit, 1994). Una vez alcanzado el primer requisito, los sistemas de CTH debían evolucionar hacia una mayor naturalidad, es decir, en la capacidad de emular la riqueza del habla humana que es intrínsecamente expresiva, ya que posee la capacidad de complementar la información verbal con una intención, actitud o estado emocional determinados. En este contexto, la mejora de la expresividad de los sistemas de CTH se ha debido a avances en el modelado de la prosodia y la generación de la señal de voz de una alta calidad.

La estructura interna de un sistema de CTH es modular y, en general, sigue las etapas mostradas en el esquema de la figura 2.3, que representa un sistema de síntesis por concatenación de unidades. En dicho esquema, en primer lugar, hay que diferenciar claramente dos procesos: la generación del corpus de voz (proceso *off-line*) y la CTH propiamente (proceso *on-line*). En segundo lugar, tal y como señala Dutoit (1997), el proceso *on-line* tiene dos módulos principales que abarcan diferentes tareas encadenadas: el procesamiento del lenguaje natural y el procesamiento digital de la señal).

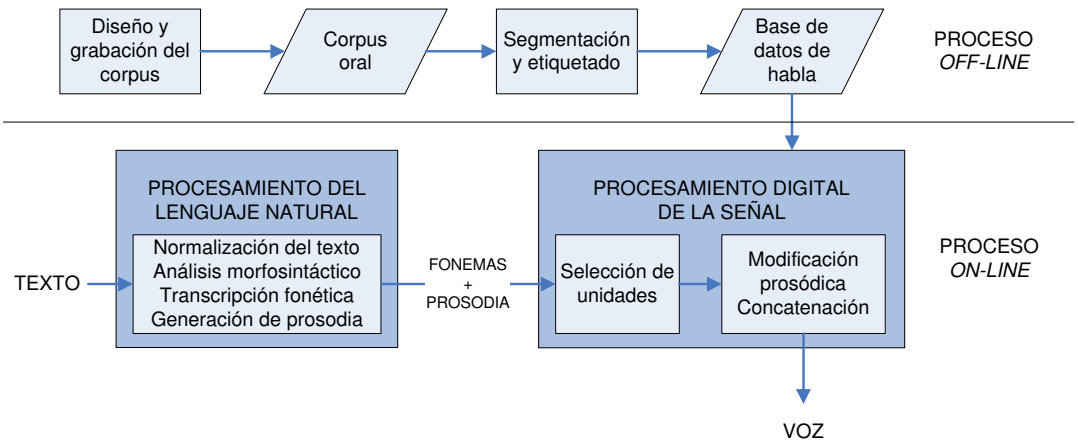


Figura 2.3: Diagrama de bloques de los dos procesos que forman parte de un sistema de CTH.

2.3.1. Procesamiento del lenguaje natural

El módulo de procesamiento del lenguaje natural (PLN) tiene como objetivo principal generar la información necesaria que dé respuesta a las preguntas: ¿qué sonidos (fonemas) se corresponden con el texto? y ¿cómo debe ser su realización sonora (prosodia)? Con esta información —segmental y suprasegmental— de entrada, el módulo de procesamiento digital de la señal sintetizará la señal de voz con la finalidad de conseguir la máxima calidad.

En este apartado únicamente se realizará una descripción superficial del cometido de las diferentes tareas que componen el módulo de PLN. Para profundizar en el análisis lingüístico orientado a la CTH se recomienda la lectura de Bonafonte et al. (2006), Llisterri et al. (2004), Montero (2003) y Dutoit (1997). A continuación se describen las cuatro tareas principales de dicho módulo:

1. La primera tarea que debe resolver el módulo de PLN es la **normalización del texto** de entrada. Generalmente, los textos presentan una serie de elementos que no son convertibles directamente en una cadena de fonemas. Dichos elementos son cifras, ordinales, horas, fechas, unidades de medida, siglas, abreviaturas, acrónimos, números romanos o símbolos especiales. Por lo tanto, se requiere de un módulo que transforme estos elementos en un texto legible. También se ocupa esta tarea del tratamiento de los signos de puntuación.

2. La segunda tarea está orientada hacia el **análisis del texto** a través de diferentes herramientas que abordan distintos niveles. Este análisis tiene que posibilitar la fijación de pausas no marcadas por los signos de puntuación, la asignación del acento y la generación de los patrones melódicos. Un análisis morfológico y una asignación de partes de la oración (POS, *part-of-speech tagging*) obtiene la categoría léxica (nombre, verbo, adjetivo, etc.) y reconoce su estructura interna mediante un análisis en morfemas. La inclusión de esta información permite mejorar la transcripción fonética, por ejemplo en palabras derivadas por prefijación, y también la asignación prosódica en palabras homógrafas. Un análisis sintáctico permite, además, estudiar las relaciones entre las palabras del texto y, por lo tanto, mejorar la asignación de la prosodia. No obstante, la incorporación de analizadores semánticos y pragmáticos que ayuden a determinar aspectos como el foco o la intención comunicativa del texto es poco frecuente.

3. La tercera tarea consiste en la **transcripción fonética** del texto normalizado. La salida de este módulo será una cadena de símbolos fonéticos que además incluyan información sobre la acentuación. La estrategia de la conversión grafema a fonema —o, si es el caso, a alófono— se basa en reglas o en diccionario (o combinación de ambos) según las características del idioma. Los sistemas basados en reglas suelen disponer de un diccionario de excepciones, mientras que los basados en diccionario necesitan de un análisis morfológico. La representación de la transcripción fonética puede seguir el Alfabeto Fonético Internacional (AFI), aunque es más habitual la uti-

lización del *Speech Assessment Methods Phonetic Alphabet* (SAMPA)[6] ya que facilita su utilización en un programa informático. Cada idioma tiene su propia representación (véase la versión para el español en el apéndice B). El principal problema que presenta SAMPA es el conflicto entre tablas para diferentes idiomas. El desarrollo del *eXtended Speech Assessment Methods Phonetic Alphabet* (X-SAMPA) ha proporcionado una tabla única de símbolos de transcripción sin diferencias específicas entre idiomas.

4. La última tarea consiste en la **generación de los rasgos prosódicos** asociados al texto con el fin de obtener una lectura natural. Los parámetros que determinan la prosodia de un enunciado son, esencialmente, la duración e intensidad segmental, la posición y duración de las pausas y el contorno de F_0 (Llisterri et al., 2003). En el ámbito de los sistemas de CTH, la bibliografía sobre el modelado prosódico es muy extensa. La curva de entonación es el parámetro prosódico más tratado por la comunidad científica, distinguiéndose entre métodos cuantitativos tales como el propuesto por Fujisaki et al. (1994), el modelo TILT (Taylor, 2000) o el basado en curvas de Bezier (Escudero, 2003) y métodos cualitativos como ToBI (Silverman et al., 1992) o Intsint (Hirst et al., 1994). Para el modelado de la duración segmental se han utilizado métodos basados en reglas (Klatt, 1979) y métodos estadísticos tales como redes neuronales (Campbell, 1990) o árboles de clasificación y regresión (Möbius y van Santen, 1996). El modelado de la intensidad es el menos presente en la bibliografía aunque se han realizado algunos trabajos específicos en esta dirección tales como los propuestos por Blecua y Acín (1995) o por Trouvain et al. (1998).

   En el habla natural, la duración de los sonidos depende del contexto en el que se encuentran. La mayoría de estudios (p.ej. Febrer et al., 1998a; Navas et al., 2002; Teixeira y Freitas, 2003), utilizan el fonema como unidad básica para la duración, aunque haya aproximaciones basadas en difonemas o sílabas. Según estos estudios, los factores que influyen en la duración de los sonidos se deben básicamente a: *i)* la identidad del fonema y los de su contexto (habitualmente, el anterior y el posterior), *ii)* información sobre el acento, *iii)* información sobre la posición del fonema en la frase y en la sílaba. Cada estudio presenta su manera particular de codificar y tratar esta información.

   La predicción de la curva de intensidad se suele llevar a cabo generalmente a nivel segmental. Aunque muchos sistemas de CTH no consideran este rasgo, los factores que deben tenerse en cuenta (Llisterri et al., 2003) están también relacionados con la identidad del segmento, el acento y la posición.

   Escudero (2003) revisa diversos trabajos relativos a las unidades de entonación para el español y los factores que caracterizan cada una de estas unidades. Existen diferentes tipos de unidades utilizadas para modelar el contorno de entonación: las unidades inferiores a la sílaba (p.ej. fonemas) y la sílaba (microentonación), el grupo acentual (GA) —relacionado con el ritmo del habla—, el grupo de entonación (GE) y otras unidades superiores (planificación del discurso). Dicho autor propone el uso del GA como unidad básica para el modelado de la entonación. Además, se concluye que algunos de los factores que deben considerarse en el nivel del grupo

[6] `www.phon.ucl.ac.uk/home/sampa`

acentual para modelar la entonación están relacionados con: *i)* el tipo de GE al que pertenece el GA, *ii)* la posición del GA dentro del GE, *iii)* la posición del acento, *iv)* la posición del GE dentro de la frase, y *v)* el número de sílabas del GA y del GE. La curva de entonación de cada GA se puede modelar con diferentes funciones (p.ej. polinomios o funciones de Bezier) o mediante una serie de puntos que estilizan el contorno (Garrido, 2001). Si el lector desea profundizar en el estudio de la entonación, especialmente para el español, puede consultar Garrido (1996).

2.3.2. Módulo de síntesis de la señal de voz

A lo largo de los últimos años han aparecido distintas técnicas para generar la señal de voz resultante en el proceso de CTH. Hay diferentes formas de clasificar estas técnicas que, a su vez, comportan una clasificación de los sistemas de CTH. Por ejemplo, Dutoit (1997) distingue claramente los sistemas basados en reglas de los que emplean concatenación de unidades. Una clasificación equivalente es la que presenta Toda (2003), aunque se refiere al segundo tipo como síntesis basada en corpus, dado que se trata de una clasificación más reciente y posterior a la aparición de sistemas que utilizan grandes corpus de voz y procesos estadísticos.

Los sistemas basados en reglas presentan una calidad poco natural debido a que la señal de voz se genera utilizando un modelo de producción del habla inexacto, ya que requiere de ciertas aproximaciones. Se recomienda la lectura de Mattingly (1974) si el lector desea tener una visión histórica de la síntesis del habla y, en concreto, si desea profundizar en los inicios y los primeros logros de la síntesis basada en reglas. Otra revisión posterior muy interesante es la presentada por Klatt (1987), en la cual el autor profundiza en este tipo de síntesis, distinguiendo los avances en la síntesis por formantes, la síntesis articulatoria y los primeros pasos hacia la síntesis por concatenación de difonemas. Ambas revisiones están accesibles en línea en la web del *Smithsonian Speech Synthesis History Project*[7].

En la actualidad, la técnica predominante en el ámbito de los sistemas de síntesis es la basada en corpus o selección de unidades (Eide et al., 2003). Estos sistemas de CTH son capaces de generar un mensaje mediante voz sintetizada, consiguiendo una buena calidad e inteligibilidad en aplicaciones de propósito general. Sin embargo, todavía se está lejos de lograr sintetizadores de habla capaces de emular toda la complejidad de la comunicación humana (Black, 2002; Schröder, 2004).

Uno de los elementos fundamentales de este tipo de sistemas es el corpus de habla, cuyo diseño influirá decisivamente en la calidad de la voz generada. En los sistemas de propósito general, éste suele estar diseñado para asegurar que la voz grabada no exhiba ningún estilo en particular, es decir, que tenga un estilo de locución neutro (Breen y Jackson, 1998). Dado que este tipo de CTH refleja claramente el estilo y la cobertura de la voz grabada (Black, 2002), la calidad del habla sintética puede variar en función de la coincidencia del texto de entrada con el contenido del corpus de propósito general diseñado. En

[7] http://www.mindspring.com/∼ssshp/ssshp_cd/ss_home.htm

cambio, los sistemas de dominio limitado suelen desarrollarse para aplicaciones específicas. En este tipo de sistemas, la calidad del habla sintetizada es muy alta cuando el texto de entrada pertenece al mismo dominio que el corpus (véase la revisión presentada en Möbius (2000)).

Por lo tanto, dada la gran influencia del contenido del corpus en la calidad del habla sintética generada, el dominio de síntesis deseado debería incluirse en el corpus. Por ejemplo, en trabajos precedentes, se han reunido diferentes emociones (Iida et al., 2003), distintos estilos de locución (Alías et al., 2004b), etc., en un mismo corpus oral. Asimismo, se hace necesario disponer en estos casos de algún método que indique el dominio más adecuado sobre el que llevar a cabo el proceso de selección de unidades.

Otro aspecto interesante es el de la flexibilidad de poder sintetizar la voz de diferentes hablantes mediante técnicas de conversión de voz (Toda, 2003). Por el momento, los sistemas que alcanzan mayor flexibilidad en este sentido son los basados en Modelos ocultos de Markov (HMM) (Yoshimura et al., 1999), que utilizan una representación estadística de los parámetros del habla junto a alguna técnica de análisis/síntesis generalmente inspirada en *vocoders*[8]. El principal reto de estos sistemas es mejorar la naturalidad del habla sintética generada sin necesidad de recurrir a un corpus de voz de gran tamaño. Recientemente, este tipo de sistemas está en auge debido a las ventajas que presentan (Black et al., 2007):

- Facilidad para modificar las características de la voz.
- Síntesis de diferentes estilos o emociones.
- Fácil aplicación a varios idiomas con pequeñas modificaciones.
- Aprovechamiento de las técnicas de reconocimiento automático del habla ya desarrolladas.
- Reducido espacio de disco duro o memoria de datos.

Con respecto a esta técnica de síntesis, el proyecto *HMM-based Speech Synthesis System* (HTS)[9] es el que mayor empuje está teniendo. Dicho proyecto dispone de una amplia recopilación de publicaciones y la posibilidad de descargar el *software* del núcleo del sistema, así como algunas voces para los idiomas inglés y japonés.

[8] acrónimo derivado del inglés *voice coder*, codificador de voz
[9] http://hts.sp.nitech.ac.jp/

Capítulo 3

Estado de la cuestión

Este capítulo describe el estado de la cuestión de los dos ejes principales que tienen relación con el presente trabajo de investigación: *i)* Las bases de datos o corpus orales para el estudio y el desarrollo de aplicaciones relacionadas con el habla expresiva (apartado 3.1) y *ii)* los dos elementos que intervienen principalmente en la síntesis del habla expresiva (apartado 3.2), que son el modelado prosódico y los métodos de síntesis que se pueden aplicar.

3.1. Corpus orales para el estudio del habla emocional

3.1.1. Introducción

En el ámbito de la interacción persona-máquina, se observa una mayor tendencia hacia el uso de la voz por parte de los usuarios, por ejemplo para consultar cierta información o realizar una determinada gestión. También se tiende a que las máquinas hablen en lugar de personas (automatización de servicios o ayuda a discapacitados). La incorporación del reconocimiento de estados emocionales o la síntesis de habla emocional pueden favorecer la comunicación haciéndola más natural (Campbell, 2000). Por lo tanto, uno de los retos más importantes en el estudio del habla expresiva es el desarrollo de corpus orales con un contenido emocional auténtico que posibiliten un análisis robusto. Este análisis tiene que proporcionar la información necesaria para abordar la tarea para la que se ha desarrollado el corpus.

Aunque en Campbell (2005) se matiza la diferencia entre corpus y base de datos, en el presente trabajo utilizaré indistintamente ambos términos cuando me refiera a conjuntos de locuciones para su utilización en alguna aplicación relacionada con las tecnologías del habla. Según dicho autor, las diferencias más importantes entre ambos conceptos radican en el diseño, el tamaño y la finalidad del conjunto de datos. Mientras la base de datos está controlada, es decir diseñada y construida para contener unos elementos concretos y, además, presenta un tamaño relativamente pequeño o limitado, el corpus es una colección

de muestras de ocurrencias naturales con un tamaño suficiente para ser representativo de los patrones que se deseen extraer a partir de su análisis.

A modo orientativo, se muestran las definiciones del *Diccionario de la lengua española* (vigésima segunda edición) de la Real Academia Española[1]:

Corpus. *Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación.*

Base de datos. *Conjunto de datos organizado de tal modo que permita obtener con rapidez diversos tipos de información.*

Como se verá más adelante, el amplio conjunto de muestras de voz que se ha recogido en el ámbito de esta tesis tiene características de corpus y de base de datos a la vez. Para ser estrictos, podríamos considerar, por un lado, que la fase de diseño y su utilización posterior —una vez segmentada y etiquetada— son propias de una base de datos y, por otro, que el tamaño y la expresividad de los textos leídos por el locutor son propios de un corpus.

3.1.2. Características principales

Según Douglas-Cowie et al. (2003), en el desarrollo de una base de datos deben considerarse cuatro aspectos principales:

- El **ámbito** que cubre una base de datos según el número de locutores, el idioma, los dialectos, el sexo de los hablantes y los tipos de estados emocionales. Estas variables son potencialmente importantes en el intento de generalización, ya que los resultados del estudio del habla y de la emoción no siempre son consistentes entre individuos, situaciones o culturas. El resumen de trabajos sobre las relaciones entre el habla y la emoción presentado en Cowie et al. (2001) muestra que algunas características del habla son consistentes entre estudios y, en cambio, otras presentan ciertas diferencias.

  La importancia de una mayor o menor variedad de locutores dependerá del objetivo de la investigación. Para propósitos de síntesis del habla, puede ser suficiente el estudio de un único hablante, de forma que su manera de expresar las emociones se modele y utilice en el proceso de síntesis. En cambio, para lograr el reconocimiento de la emoción a través del habla, se necesitan bases de datos que contengan la máxima variedad de signos por los cuáles una emoción dada se pueda expresar.

  El otro aspecto fundamental es el rango de emociones que pretende cubrir la base de datos. Muchos trabajos se decantan por las emociones básicas o plenas, aunque no hay un consenso sobre cuales forman este conjunto (véase la tabla 2.1). Otros investigadores defienden la idea de que es más práctico el uso de estados relacionados con la emoción (véase el apartado 2.1), ya que son mucho más frecuentes que las emociones plenas.

[1] http://www.rae.es/

- La **naturalidad** de las locuciones del corpus que dependerá del modo en que se han llevado a cabo las grabaciones. Inicialmente, hay que diferenciar entre voz de actor o actriz y voz natural (tanto si se trata de producciones espontáneas como inducidas). La voz de actor puede obtenerse mediante una interpretación o la lectura de un texto. El debate se centra en el compromiso entre la autenticidad de la emoción expresada y el control sobre la grabación. La decisión de utilizar un tipo de corpus u otro dependerá también del objetivo de la investigación o la tarea que se pretenda desarrollar.

- El **contexto** en el cual se produce una locución que complementa claramente el significado emocional percibido por el oyente. Es importante que las bases de datos representen dicho contexto, ya que la expresión de la emoción se puede deber básicamente a su presencia más que al contenido vocal. Se distinguen cuatro formas básicas de contexto: el semántico (palabras concretas con un elevado contenido emocional), el estructural (patrones de entonación, énfasis, ritmo del habla), el intermodal (expresión facial, gestos y postura) y el temporal (presencia de cambios acústicos en determinados momentos del discurso).

- Los **descriptores** que permiten representar tanto el contenido lingüístico como emocional y los parámetros acústicos del habla. El etiquetado del contenido emocional está relacionado con la naturalidad, ya que el material grabado por un actor puede clasificarse fácilmente. En cambio, el habla espontánea puede presentar un amplio abanico de estados emocionales difíciles de clasificar y de tratar estadísticamente. Los descriptores del habla deben cubrir todo el rango de características relacionadas con la expresión vocal de la emoción (cualidad de la voz, prosodia y elementos no lingüísticos, como la risa o el llanto). En función del parámetro se puede escoger una representación cualitativa (o en categorías) o cuantitativa.

3.1.3. Recopilaciones de corpus de habla emocionada

No es el objetivo del presente trabajo realizar un compendio exhaustivo de las bases de datos disponibles para el estudio del habla emocionada, ya que recientemente han aparecido diversos estudios sobre habla y emoción. En Douglas-Cowie et al. (2003) se ofrece una recopilación de 21 bases de datos con una descripción del ámbito (número de sujetos, descripción de las emociones, idioma), la naturalidad (simulada, semi-natural o natural y si está transcrita) y el contexto (temporal y modo). Respecto al modo, sólo se distingue entre audio y audiovisual. En Cowie et al. (2005) se presenta una nueva recopilación con 48 bases de datos, en la que se observa un incremento notable de bases de datos multimodales y distinguiéndose hasta 4 modos diferentes: audio, vídeo de la cara, gestos y medidas fisiológicas. En Ververidis y Kotropoulos (2003) se revisan 32 bases de datos de habla emocional, proporcionando una descripción básica de cada una y su aplicación. Este conjunto de bases de datos se amplía a 64 en una revisión posterior (Ververidis y Kotropoulos, 2006).

3.1.4. Clasificación según la estrategia de grabación del corpus

Según la estrategia que se haya seguido para conseguir una base de datos de habla emocionada se puede establecer una primera clasificación. De los aspectos descritos en el apartado 3.1.2, la naturalidad es el más afectado por el modo de conseguir voz emocionada. A continuación, se describen las cuatro categorías propuestas por Campbell (2000) y seguidas por otros autores (Schröder, 2004).

3.1.4.1. Habla natural recopilada

La interacción humana espontánea es la que presenta un habla emocional con mayor naturalidad y, por lo tanto, un corpus formado por grabaciones de este tipo proporcionaría los datos más apropiados para el análisis. La justificación para el uso de este tipo de material es la pureza de las emociones del sujeto al cual pertenece el contenido oral. Sin embargo, conseguir un contenido emocional real presenta problemas en distintas direcciones: la falta de control sobre el contenido, la calidad de sonido, la dificultad del etiquetado de los estados emocionales y por último, los aspectos legales y éticos que pueda conllevar.

Tal estilo de habla es difícil de obtener, ya que no se han desarrollado las herramientas necesarias para poder tratarlo de forma robusta. Una fuente de este tipo de material son ciertos programas de radio y televisión, aunque el requerimiento de una adquisición de sonido de alta calidad a menudo no es posible debido a la carencia de herramientas capaces de tratar con la variación de la distancia entre el locutor y el micrófono, la reverberación, el ruido o la superposición de voces. De forma similar, los humanos podemos reconocer expresiones emocionales en caras con un cierto grado de rotación o a cierta distancia. En cambio, esta no es una tarea fácil para las técnicas de modelado de la imagen, ya que son poco robustas ante variaciones en la iluminación, la aparición de oclusiones o rotaciones (Melenchón, 2006). Además, los aspectos legales referentes al uso público de este tipo de material limitan la libertad de los investigadores y desaniman a los propietarios para ponerlos a disposición pública (Campbell, 2000).

Algunos ejemplos de las bases de datos naturales más significativas son *The Reading-Leeds database*, *The Belfast Naturalistic database* y la base de datos *CREST*, descritas sucintamente en Douglas-Cowie et al. (2003).

3.1.4.2. Habla emocional inducida

Provocar emociones auténticas en personas en un laboratorio es una forma de compensar algunos de los problemas descritos anteriormente. De todas formas, no se trata de una tarea fácil ya que, en un entorno seguro y controlado como es el laboratorio, las emociones extremas quedan fuera de lugar y el hecho de inducir deliberadamente emociones como el miedo o el enfado es éticamente cuestionable (Campbell, 2000). En Schröder (2004) se describen cinco tipos de procedimientos de inducción del estado de ánimo, aunque existen pocas bases de datos de este tipo orientadas al estudio del habla emocional.

3.1.4.3. Habla emocional estimulada

Este método consiste en la lectura de textos con un contenido verbal apropiado para la emoción que se quiere expresar. En Campbell (2000), para validar si este tipo de habla era capaz de evocar emociones genuinas, se realizó un test de percepción con locuciones generadas mediante síntesis concatenativa de textos semánticamente neutros, pero manteniéndose la prosodia y la cualidad de la voz de los originales. Los resultados obtenidos en este test mostraron un alto porcentaje de identificación correcta.

Una idea similar se siguió en la creación de la *Belfast Structured Emotion Database* (Douglas-Cowie et al., 2003) en la cual 38 locutores leyeron dos párrafos interpretando cada una de las cuatro emociones básicas (enfado, miedo, tristeza y alegría) y un estilo neutro como referencia.

Una desventaja de este método radica en la dificultad de comparar frases con diferentes textos. Esta inevitable pérdida en el control sobre el contenido de las frases debe contrarrestarse con un incremento del número de frases para que métodos estadísticos permitan generalizar modelos.

3.1.4.4. Habla emocional de actor

La mayor parte de bases de datos de habla emocionada se han grabado con actores. La gran ventaja de este método es el control sobre el contenido verbal y fonético del habla, ya que todos los estados emocionales se pueden emular utilizando las mismas frases. Esta estrategia permite comparaciones directas de los aspectos segmentales, la prosodia y los parámetros asociados con la cualidad de la voz para los diferentes estados emocionales expresados. Además, existe la posibilidad de obtener expresiones correspondientes a emociones plenas (Schröder, 2004).

El gran inconveniente que presenta esta estrategia es que no asegura que las locuciones obtenidas representen plenamente las características del habla utilizada por las personas cuando, de forma natural, experimentan emociones similares (Campbell, 2000). Según el mismo autor, otro aspecto a tener en cuenta es que en la interacción social habitual se ha constatado una tendencia a disimular los sentimientos personales mediante el control de los elementos expresivos. También se da la circunstancia de expresar emociones que no son sentidas. Puede darse el caso de que los oyentes reconozcan una emoción intencionada en la voz de un actor, pero que no sea sentida o sincera, ya que se corre el riesgo de que se reproduzcan formas estereotipadas de una emoción concreta.

3.1.5. Clasificación según la aplicación

Los diferentes enfoques en la investigación sobre habla y emoción están muy relacionados con las tareas en las que se pretende aplicar dicha investigación. Por lo tanto, para una determinada investigación se requerirá una base de datos con las características

(ámbito, contexto, naturalidad y descriptores) adecuadas. Hay que distinguir claramente entre los procesos de expresión o percepción. En Schröder (2004) se presenta una adaptación del modelo de lentes de Brunswik propuesta por Klaus Scherer para ilustrar el proceso de inferencia de emociones entre dos personas (véase la figura 3.1). La investigación en este ámbito se puede subdividir según sea el elemento central del modelo, pudiéndose distinguir entre *estudios centrados en el hablante*, *estudios de codificación* o *estudios centrados en el oyente*.

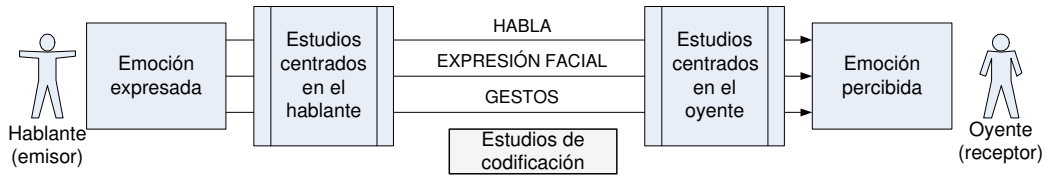


Figura 3.1: Tipos de estudios sobre habla y emoción según el elemento central

3.1.5.1. Reconocimiento de la emoción en el habla

El objetivo de los estudios *centrados en el hablante* (Schröder, 2004) es establecer la relación entre el estado emocional del hablante y un conjunto de parámetros cuantificables del habla. Se trata, generalmente, del reconocimiento de emociones de un hablante a partir de la señal de voz. Según Devillers et al. (2005), uno de los retos en el análisis del habla real es la identificación de indicadores orales atribuibles al comportamiento emocional y que no sean simplemente características propias del habla conversacional espontánea. En dicho trabajo se muestra una tabla que resume 14 artículos sobre experimentos de detección automática de la emoción, en la que se indica el estilo del corpus, el tamaño, las etiquetas emocionales, las características del habla, el método de aprendizaje automático y la tasa de detección. Existen muchas características del habla relacionadas con la emoción como las prosódicas, las espectrales y las de cualidad de la voz. Además, la información léxica, la detección de disfluencias del habla (dificultad en el flujo normal del habla) o la presencia de sonidos no verbales como la risa, pueden ser útiles para la detección de la emoción.

En Ververidis y Kotropoulos (2006) aparecen 47 bases de datos orientadas al reconocimiento automático de las emociones en una recopilación que contiene un total de 64 bases de datos de habla emocional, en la que se indican el número de estados emocionales, el idioma, el número de hablantes y la estrategia de grabación. Además, se presentan las características acústicas utilizadas mayoritariamente para el reconocimiento del habla emocionada y las técnicas apropiadas para clasificar el habla en estados emocionales.

3.1.5.2. Codificación de indicadores de la emoción

Podemos encontrar un conjunto de bases de datos orientadas al estudio de aquellos parámetros acústicos y la correlación entre ellos que sean relevantes en la percepción de la

emoción. Se trata de grandes bases de datos de habla emocionada genuina y natural que intentan abarcar un amplio rango de emociones (Douglas-Cowie et al., 2003). El objetivo inicial de estas bases de datos es el desarrollo de sistemas completos de anotación que recojan el contenido emocional percibido junto con una descripción lingüística (p.e. transcripción ortográfica, entonación y otros efectos prosódicos y paralingüísticos). En etapas posteriores, dichas bases de datos se pueden utilizar en aplicaciones de reconocimiento de emociones (*The Belfast Naturalistic database*) o en síntesis de habla emocionada (CREST).

De este conjunto cabe destacar la base de datos del proyecto JST/CREST ESP[2], cuyo objetivo es el desarrollo de interfaces avanzadas para la interacción mediante lengua hablada. La base de este proyecto es la producción y el análisis de un enorme corpus de interacciones orales cotidianas. Durante cinco años se ha llevado a cabo la grabación de la voz de un pequeño conjunto de locutores voluntarios en situaciones ordinarias de la vida. En Campbell (2002, 2004) se puede ampliar la información sobre los aspectos técnicos de las grabaciones, los locutores y el proceso de anotación de este corpus. Cabe destacar que el etiquetado de la emoción ha revelado la existencia de muy pocas emociones plenas y el predominio de un contenido emocional medianamente positivo. Por lo tanto, este corpus destaca más por una amplia variación en la dimensión social derivada de la interacción ordinaria de los locutores que por el contenido de emociones plenas.

3.1.5.3. Síntesis del habla emocionada

Según Schröder (2004), los estudios *centrados en el oyente* modelan los parámetros del habla con el objetivo de transmitir un cierto estado emocional. El efecto perceptivo asociado a un cierto estímulo controlado ha sido objeto de numerosos estudios en este ámbito (p.ej. Montero et al. (1999a)). El tipo de descripción de los estados emocionales utilizado tiene un papel muy importante en los resultados obtenidos. Otro aspecto importante es la elección de los parámetros del habla que se van a modificar para intentar simular emociones.

Existen diferentes recopilaciones de bases de datos orientadas a la síntesis del habla emocional (Murray y Arnott, 1993; Douglas-Cowie et al., 2003; Ververidis y Kotropoulos, 2003; Schröder, 2004), aunque el reciente resumen presentado en Ververidis y Kotropoulos (2006) se puede destacar como uno de los más completos, ya que menciona un total de 16 bases de datos de este tipo.

Las bases de datos orientadas a la síntesis del habla emocional suelen tener las siguientes características:

- La estrategia de grabación suele consistir en utilizar un actor o un locutor profesional que lea un conjunto de textos con las emociones que se quieren simular. Existen dos posibilidades en cuanto a la naturaleza de los textos: *i)* conjunto de textos neutros —sin contenido emocional— que se repiten para cada emoción, o *ii)* textos con contenido emocional. Mientras que el primer tipo de texto facilita la comparación

[2] `http://feast.his.atr.jp`

entre estilos, ya que el contenido es el mismo para todas las emociones, el segundo facilita la simulación de la emoción por parte del actor o locutor.

- La duración del corpus suele ser de varias horas, especialmente en el caso de que la síntesis sea basada en corpus. En este tipo de síntesis se requieren diferentes subcorpus que contengan los diferentes estilos y que bien pueden utilizarse de forma independiente (*tiering*) bien pueden mezclarse (*blending*) permitiendo cambios graduales entre tipos de voz y estilos mezclados (Black, 2003).

3.1.6. Corpus de habla emocional en la investigación de ámbito nacional

A continuación se describen brevemente algunos de los principales corpus de habla emocional desarrollados en el ámbito nacional que se han aplicado principalmente al modelado y a la síntesis del habla emocional.

3.1.6.1. Spanish Emotional Speech (SES)

La base de datos *Spanish Emotional Speech* (SES) fue desarrollada por el Grupo de Tecnología del Habla (GTH) de la Universidad Politécnica de Madrid en el marco del proyecto VAESS (Montero et al., 1998; Montero, 2003). La grabación se llevó a cabo con un actor profesional en dos sesiones de estudio. Consta de 4 emociones (tristeza, alegría, enfado y sorpresa) y un estilo neutro. Respecto al diseño de los textos, estos se dividieron en tres tipos:

- 15 frases cortas de carácter neutro, sin ninguna emotividad, buscando un equilibrio fonético.
- 31 palabras aisladas extraídas de las frases anteriores manteniendo, sin embargo, algunos grupos acentuales enteros.
- 3 párrafos, también de carácter neutro, y un cuarto párrafo formado por doce de las anteriores frases con una cierta estructura narrativa que permitió comparaciones en tres contextos diferentes.

El etiquetado y marcado del corpus se realizó de forma semimanual utilizando técnicas de resíntesis para detectar posibles errores y refinar el etiquetado.

Se llevó a cabo una prueba de escucha para evaluar la identificación de emociones con la voz grabada por el actor. Se obtuvieron unos resultados de identificación correcta de prácticamente el 90 %, excepto para la alegría, que obtuvo un 74 %.

Además, Montero (2003) detalla un análisis cualitativo de las 4 emociones y un análisis cuantitativo de las duraciones, el ritmo y la entonación comparándolas con el estilo neutro.

3.1.6.2. Interface Emotional Speech Synthesis Database (IESSDB)

El *Grup de Tractament de la Parla* de la *Universitat Politècnica de Catalunya* participó en la producción de la base de datos IESSDB (Hozjan et al., 2002). Esta base de datos se grabó en cuatro idiomas —francés, inglés, esloveno y español—, y contiene frases de dos actores profesionales simulando cada uno de los seis estados emocionales ("Big Six") adoptados por el estándar MPEG-4 (enfado, asco, miedo, alegría, tristeza y sorpresa) más un estilo neutro suplementario. Su diseño se orientó principalmente a la síntesis del habla, aunque también se ha utilizado en experimentos de reconocimiento de emociones en el habla (Nogueiras et al., 2001). La versión para el español consta de 100 frases enunciativas, 34 frases interrogativas, 16 párrafos y 34 palabras aisladas, de las que 10 se corresponden a los dígitos.

La evaluación subjetiva del contenido emocional se realizó mediante una prueba de identificación con 16 oyentes a los que se les permitió marcar dos opciones para cada estímulo. Los resultados obtenidos fueron bastante satisfactorios (cerca del 80 % de identificación contando únicamente la primera opción y un 90 % si además se consideraba como correcta la segunda opción marcada por los participantes en la prueba).

3.1.6.3. AHOLAB

El grupo AhoLab del Departamento de Electrónica y Telecomunicaciones de la Universidad del País Vasco ha desarrollado una base de datos de habla emocionada en euskera, de tamaño medio y formada por dos subcorpus que se distinguen por el contenido semántico de los textos (Navas et al., 2006). Se trata de su primera aproximación a la generación de un corpus oral que contenga locuciones de las emociones denominadas "Big Six" descritas en el apartado 2.1.3.1. Este corpus inicial se ha dividido en un subcorpus que incluye textos con un contenido semántico neutro e igual para todas las emociones (Subcorpus común), y otro que incorpora textos semánticamente relacionados con cada emoción (Subcorpus específico). El objetivo de este estudio fue valorar si un corpus textual con contenido neutro era suficiente para generar una base de datos de habla emocional para su posterior utilización en síntesis basada en corpus. Ambos subcorpus se diseñaron para incluir palabras aisladas y frases tanto enunciativas como interrogativas. Se utilizó una actriz de doblaje profesional para la grabación. La base de datos constó de 1h y 25 min de voz grabada, de la que 50 minutos corresponden al subcorpus común (además de las 6 emociones se grabó un estilo neutro) y 35 minutos al subcorpus específico.

La evaluación subjetiva del contenido emocional de ambos subcorpus se realizó mediante dos pruebas, una con vascohablantes y otra con personas que no entendían el euskera. Los resultados de la identificación global mostraron una tasa de reconocimiento promedio mayor para el subcorpus específico que para el común (véase la tabla 3.1). Como era de esperar, las tasas con un mayor reconocimiento fueron para el grupo de evaluadores vascohablantes en el subcorpus específico, lo que indica que el contenido semántico puede contribuir a la identificación de la emoción.

Tabla 3.1: Tasas de reconocimiento de las pruebas subjetivas (Tabla extraída de Navas et al., 2006)

	Subcorpus común	Subcorpus específico	Total
Vascohablantes	65.9 %	85.8 %	76.2 %
No vascohablantes	51.3 %	46.6 %	48.9 %
Total	58.6 %	66.2 %	62.6 %

El análisis acústico de la base de datos se centró en parámetros relacionados con la frecuencia fundamental y la energía, llevándose a cabo un análisis de la varianza (ANOVA) con el objetivo de estudiar las diferencias en la distribución de esos parámetros en los dos subcorpus. Además, se calcularon las medias de todos los parámetros para cada emoción con la finalidad de ajustar globalmente la frecuencia fundamental y la energía en el sintetizador de habla emocionada.

Finalmente, se realizó un experimento de identificación automática de las emociones mediante parámetros prosódicos, cuyos resultados mostraron un alto grado de correlación con los resultados de la prueba subjetiva llevada a cabo con el subcorpus común y el grupo de evaluadores vascohablantes.

3.1.6.4. LAICOM

El Laboratorio de Análisis Instrumental de la Comunicación de la Universidad Autónoma de Barcelona (LAICOM-UAB) desarrolló un corpus oral con el objetivo de realizar una modelización acústica de la expresión emocional en el español (Rodríguez et al., 1999). El corpus de 336 discursos se construyó mediante la interpretación de dos textos por parte de 8 locutores (4 hombres y 4 mujeres) que los repitieron con tres niveles de intensidad para las 7 emociones consideradas como básicas en dicho estudio: alegría, deseo, rabia, miedo, sorpresa, tristeza y asco.

Con la finalidad de construir el corpus definitivo a partir del cual realizar el análisis acústico de las emociones, se realizó una prueba de percepción que permitiese escoger las interpretaciones que mejor representasen cada emoción. Cada discurso emocionado (con una duración de 20 a 40 segundos) fue escuchado por dos grupos de más de 30 oyentes. Cada grupo valoró 30 interpretaciones mediante tres tareas: *i)* indicar qué emoción o emociones reconocía en cada voz; *ii)* asignar un grado de verosimilitud al locutor; *iii)* especificar si se había emocionado y en qué grado. En total participaron 1.054 oyentes, en su mayoría estudiantes.

Los resultados de esta prueba permitieron decidir con objetividad qué interpretaciones contenían realmente información acústica asociada a las emociones. Se seleccionaron las 4 o 5 interpretaciones de cada emoción con mayor porcentaje de identificación y un grado de verosimilitud más alto.

El posterior análisis acústico se centró en medidas de la frecuencia fundamental, de

la presión sonora y en parámetros asociados con el ritmo. En este estudio destaca que no se grabó un estilo neutro, sino que se compararon los resultados del análisis con la media aritmética de los datos de cada parámetro para cada locutor (estado-promedio).

La experiencia obtenida con este trabajo sirvió de base para el experimento de síntesis de habla emocionada descrito en Iriondo et al. (2000) sobre el que se ofrecen más detalles en el apartado 5.1.1.

3.2. Síntesis del habla expresiva

Según Tatham y Morton (2003), se entiende por habla expresiva aquella en la que un oyente puede detectar cierta emoción, actitud o intención por encima del significado básico que transmiten las palabras del mensaje oral y la forma en que se estructuran sintácticamente. En una obra teatral, la frase *'Yo la quiero'* se debe acompañar de una acotación que permita la interpretación adecuada. Para esta frase, indicaciones tales como *'Sinceramente'*, *'Con enfado'* o *'Soñando'* serían posibles según el contexto de la escena.

En cierta manera, se puede entender la síntesis del habla expresiva como un procedimiento capaz de generar, en su salida, una señal de voz lo más natural y auténtica para una entrada con la estructura siguiente:

Locutor: [emoción] texto.

Este tipo de síntesis del habla constituye una cuestión muy compleja, ya que intervienen diferentes áreas del conocimiento. En primer lugar, la psicología trata de describir los estados emocionales y las acciones que realizan los humanos para expresarlas y percibirlas. En la bibliografía (véase el apartado 2.1.3), encontramos principalmente dos puntos de vista para describir las emociones: *i)* divididas en categorías discretas o *ii)* como puntos de un espacio multidimensional. En segundo lugar, la psicoacústica analiza el efecto de los estados emocionales en el habla (véase el apartado 2.2.1). Se parte de la hipótesis que la voz sufre cambios acústicos causados directamente por alteraciones fisiológicas cuando una persona experimenta un determinado estado emocional (Scherer, 1986). Por lo tanto, se requerirá una modelización acústica adecuada consistente en la definición y cuantificación de los parámetros del habla que están relacionados con la percepción de las emociones. En este sentido, disponer de un corpus de habla expresiva resulta indispensable para la consecución del correspondiente modelo (véase el apartado 3.1). Finalmente, la investigación en tecnología del habla tiene que basarse en todo este conocimiento para conseguir sintetizar habla expresiva, mediante la incorporación de dichos modelos y corpus. La modelización acústica obtenida reflejará las características propias del locutor o de los locutores del corpus.

El objetivo actual de mejorar la naturalidad de los sistemas de CTH, ha hecho confluir desarrollos producidos por empresas del ámbito de las tecnologías del habla, debido a la creciente demanda en sistemas interactivos, con los avances derivados de la investigación en las teorías sobre producción y percepción del habla.

En los siguientes apartados se revisan los avances más significativos en el campo de la síntesis del habla expresiva. En primer lugar, se analizan diferentes enfoques del modelado prosódico y, a continuación, se describen los métodos de síntesis de habla más habituales.

3.2.1. Modelado prosódico para la síntesis del habla expresiva

Es bien conocido el papel de la prosodia en la expresividad vocal tal y como se ha comentado previamente en el apartado 2.2.1, destacando la entonación, el ritmo y la intensidad como los principales parámetros del habla que aportan información extralingüística y paralingüística. La percepción de estos tres elementos está relacionada principalmente con:

- La evolución de la frecuencia fundamental, especialmente en lo que se refiere a los valores medios, la variación y la forma del contorno.
- La duración de los sonidos del habla y la frecuencia y la duración de las pausas.
- La amplitud de unos sonidos respecto los otros.

Los valores de estos tres parámetros dependen del contexto, pero también presentan unos valores intrínsecos (Tatham y Morton, 2003). Por ejemplo la vocal /a/ tiene una intensidad intrínseca mayor que la vocal /i/, y puede darse el caso que a veces una /i/ acentuada tenga menor intensidad que una /a/ átona. Esta combinación de valores intrínsecos y relativos, así como la interacción entre estos parámetros prosódicos y otros como los de cualidad de la voz y, por último, su función lingüística, complica el modelado individual de parámetros para emociones concretas.

En la bibliografía encontramos distintos estudios que proponen una cuantificación de los parámetros del habla, pero que están claramente condicionados por las bases de datos de voz utilizadas y por los objetivos concretos de cada estudio. Si el lector desea tener una visión global de las publicaciones que abordan el modelado prosódico mediante reglas para la síntesis de habla emocional, puede consultar Schröder (2004). En dicho resumen se muestran 11 tablas con las reglas prosódicas para las emociones más frecuentemente analizadas (alegría, tristeza, rabia, miedo, sorpresa y aburrimiento). Existen modelos para el alemán (2), el inglés americano, el castellano (2), el japonés, el holandés, el inglés británico (2), el inglés irlandés y el alemán austriaco. Los dos modelos descritos para el castellano son los presentados en Montero et al. (1999a) e Iriondo et al. (2000).

J.M. Montero, del Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica de la Universidad Politécnica de Madrid, ha centrado su tesis doctoral (Montero, 2003) en la mejora de la naturalidad de los sistemas de CTH en castellano mediante la incorporación de variedad emocional.

Para el modelado prosódico del habla con emociones se grabó el corpus SES (descrito en el apartado 3.1.6), que consistió, fundamentalmente, en la interpretación por parte de un actor profesional de 38 años de un conjunto de textos formado por frases cortas, palabras y párrafos. Se emplearon tres emociones primarias simuladas (tristeza, alegría y enfado) y, además, se grabó igualmente una emoción calificada como secundaria (sorpresa) para su utilización posterior.

Se llevó a cabo un análisis cualitativo de las emociones y otro cuantitativo de los parámetros relacionados con el ritmo y la entonación. A modo de ejemplo, se reproducen

los valores de dos de las tablas del modelo prosódico de entonación y duración obtenido a partir del análisis del corpus SES (véanse las tablas 3.2 y 3.3).

Tabla 3.2: Resultados del análisis cuantitativo de la entonación de las frases del corpus SES para las diversas emociones extraídos de Montero (2003)

Parámetro	Alegría / Neutra	Tristeza / Neutra	Sorpresa / Neutra	Enfado / Neutra
F0 de la primera tónica	1,29	0,83	1,61	0,96
Pendiente de declinación de las tónicas	1,82	0,76	-1,44	-0,05
F0 de la 1a sílaba	1,23	0,76	1,12	0,90
F0 del último valle no oxítono (enunciativa)	0,91	0,68	1,47	0,90
F0 de la última tónica no oxítona (enunciativa)	1,32	0,79	2,51	1,19
F0 del último fonema no oxítono (enunciativa)	1,07	1,00	1,78	1,25
F0 de la 1a sílaba (interrogación)	1,08	0,76	1,06	0,95
F0 del último valle (interrogativa)	1,15	0,84	1,45	1,13
F0 de la última tónica (interrogativa)	1,55	0,91	1,18	1,34
F0 del último fonema (interrogativa)	1,12	0,64	1,56	0,90

Tabla 3.3: Resultados del análisis cuantitativo de diversos parámetros de duración de las frases del corpus SES para las diversas emociones extraídos de Montero (2003)

Parámetro	Alegría / Neutra	Tristeza / Neutra	Sorpresa / Neutra	Enfado / Neutra
Efecto medio del contexto para las consonantes	0,9222	1,0607	1,0224	0,9831
Efecto medio del contexto para los diptongos	0,9627	1,1620	1,1233	1,0538
Efecto medio del contexto para las vocales	0,9969	1,1003	1,1067	0,9168
Efecto del alargamiento vocálico prepausa	0,9086	1,2816	1,0398	0,7811
Efecto medio del número de sílabas	1,0116	1,1326	1,1903	1,0961
Duración media de todos los fonemas	1,0498	1,2629	1,1464	1,2289
Duración media de las vocales	1,0664	1,0296	1,1164	1,1003
Duración media de las diptongos	0,9952	1,0736	1,0622	1,0208
Duración media de las consonantes	1,1114	1,5303	1,2516	1,4994

3.2.2. Métodos de síntesis aplicados al habla expresiva

Como se ha mencionado al inicio del presente capítulo, la síntesis del habla expresiva forma parte de los estudios centrados en el oyente; por lo tanto, el objetivo principal es que el receptor del mensaje perciba a través del habla sintetizada la emoción que se desea emular. Conseguir este objetivo implica entender y modelar la cadena entera de relaciones que se producen en el proceso comunicativo.

Uno de los sistemas pioneros en la síntesis del habla expresiva fue el *Affect Editor* (Cahn, 1989),un *software* basado en la identificación de los efectos de la emoción en el habla así como en la elección de una representación apropiada. Dicho programa implementa un modelo acústico del habla que genera las instrucciones para producir el efecto deseado. La autenticidad de la emoción estaba limitada por las capacidades del sintetizador y por una descripción incompleta de los fenómenos acústicos y perceptivos, especialmente, entre el texto y los parámetros acústicos.

Para modelar los efectos de la emoción en el habla se desarrolló un modelo acústico representado por un conjunto de parámetros, suponiendo un comportamiento indepen-

diente de cada parámetro. Se definieron 4 categorías de parámetros que variaban en una escala común de -10 a 10. Los parámetros pertenecen a cuatro categorías:

- El tono, definido como la respuesta perceptiva a la F_0.
- Los parámetros relacionados con el tiempo, que controlan la velocidad del habla y el ritmo, que está relacionado con el acento y las pausas.
- La cualidad de la voz, que se describe mediante diferentes propiedades que se pueden medir en la señal de voz como la cantidad de ruido fricativo, el brillo (relación de baja y alta frecuencia) o el *jitter*.
- La articulación entendida como el grado de precisión en la pronunciación de los fonemas.

La cuantificación de estos parámetros se tuvo que traducir a los ajustes del sintetizador utilizado. Dicho editor trabajaba con el sistema DECTalk[3] y se llevó a cabo un experimento de evaluación que dio como resultado una tasa de reconocimiento de emociones básicas (enfado, asco, alegría, tristeza, miedo y sorpresa) del 78,7 %.

Un avance posterior en este campo fue el llevado a cabo por Murray y Arnott (1995), quienes desarrollaron el sintetizador de habla emocionada HAMLET, que se basaba en reglas y que también utilizaba el sistema DECTalk. En este caso, las emociones básicas fueron también seis, coincidiendo con Cahn (1989) salvo en el asco, que fue sustituido por la pena. Los experimentos de reconocimiento de emociones realizados mostraron que el hecho de añadir voz sintética emocionada a textos emotivos ayudaba significativamente en la identificación subjetiva de la emoción pretendida. En cambio, la mejora no fue tan significativa con textos neutros. No todas las emociones obtuvieron unas tasas de reconocimiento parecidas y, además, no se mantuvo el orden las tres emociones mejor identificadas para textos emotivos y textos neutros. Con textos emotivos, las tres emociones mejor identificadas fueron en este orden enfado, pena y tristeza. En cambio, con textos neutros, la tristeza pasó a ser la primera. Las tres emociones peor identificadas fueron, en ambos casos, el asco, el miedo y la alegría. Para este prototipo se partió del conocimiento sobre la correlación entre habla y emoción (Murray y Arnott, 1993) y se realizaron ajustes mediante técnicas heurísticas con el fin de mejorar el realismo del habla sintética emocionada.

En los siguientes años siguieron apareciendo diferentes propuestas de síntesis de habla expresiva, de las cuales destaca el proyecto VAESS, en cuyo entorno se desarrolló el corpus oral SES, descrito en el apartado 3.1.6. El proyecto propició nuevos experimentos de síntesis, primero por formantes (Montero et al., 1998) y luego mediante concatenación de unidades de voz emocionada (Montero et al., 1999a,b).

En el año 2000 tuvieron lugar unas importantes jornadas sobre habla y emoción (Cowie et al., 2000b), en las que se presentaron aproximaciones basadas en formantes

[3]Versión comercial del sintetizador por formantes de Klatt desarrollado por Digital Equipment Corporation

(Burkhardt y Sendlmeier, 2000), concatenación de difonemas (Vine y Sahandi, 2000; Murray et al., 2000; Iriondo et al., 2000) y concatenación por selección de unidades (Iida et al., 2000). Se concluyó que la síntesis por concatenación presentaba mayor calidad, pero carecía de la versatilidad de la síntesis por formantes para modificar los parámetros del habla. Posteriormente, Schröder (2001) presentó una completa revisión de los sistemas desarrollados hasta ese momento en la que realiza una interesante comparación entre estos.

La actividad en el campo de la síntesis del habla expresiva ha ido en aumento hasta el momento presente, aunque todavía conviven diferentes técnicas de síntesis que ejemplificamos en los puntos siguientes:

Concatenación de difonemas. En primer lugar, el experimento llevado a cabo por Bulut et al. (2002) consistió en combinar la prosodia y las unidades de un corpus de 4 emociones (enfado, alegría, tristeza y neutro). Tras un experimento realizado con 80 frases de prueba generadas mediante concatenación de difonemas y modificación basada en TD-PSOLA, se concluyó que la mejor configuración era utilizar modelos prosódicos y corpus específicos para cada emoción. Por otra parte, Schröder y Trouvain (2003) presentaron un sistema de CTH para el alemán denominado MARY (*Modular Architecture for Research on speech sYnthesis*) que implementaba un módulo de síntesis por difonemas con la técnica *MultiBand Resynthesis Over-Lap Add* (MBROLA). Se implementó un primer sintetizador de habla emocionada con una representación en un espacio bidimensional (activación y evaluación) según un conjunto inicial de reglas prosódicas basadas en Schröder et al. (2001). Se constató la necesidad de una mayor versatilidad del sintetizador en cuanto al control de parámetros relacionados con la cualidad de la voz y del contorno de la F_0. Un tercer y último ejemplo de este tipo es un sistema de generación de habla expresiva para la narración de historias en holandés (Theune et al., 2006). La utilización de un sistema de CTH comercial basado en difonemas condicionó el experimento, ya que solo se logró evaluar el modelado prosódico y, al igual que en el sistema MARY, se detectó la necesidad de mayor expresividad vocal por parte del sintetizador.

Selección de unidades. Mediante esta técnica de síntesis, Iida et al. (2003) presentaron un sistema de síntesis con tres emociones (alegría, tristeza y enfado) en el que se utilizó CHATR (Black y Taylor, 1994). Se grabó un corpus recurriendo a un locutor y a una locutora no profesionales, aunque con cierta experiencia en expresión vocal. Cada locutor leyó, con el estilo adecuado, un conjunto de textos emotivos. La duración de cada subcorpus osciló entre los 30 y 60 minutos. La síntesis de cada emoción se llevó a cabo de forma independiente. Los resultados de la prueba subjetiva realizada mostraron ciertas confusiones, sobre todo de los estímulos alegres que se confundieron mayoritariamente por neutros. Finalmente, propusieron estudiar nuevos correlatos del habla con determinadas emociones, tanto en el nivel segmental como en el suprasegmental, para poder enriquecer el criterio de selección de las unidades. Por otra parte, Black (2003) argumenta la necesidad de usar múltiples estilos para algunas aplicaciones de síntesis del habla, aunque señala que raramente se requieren emociones plenas. La síntesis por selección de unidades es una buena opción si se dispone del estilo requerido, pero conlleva el problema del aumento de voz grabada al ir ampliando estilos. Black constata también que el aburrimiento que

producen la mayor parte de los sistemas de CTH se puede deber al modo utilizado para grabar la voz (muchas horas de voz en un estudio profesional). Más recientemente, y con la intención de mejorar la naturalidad del habla, se ha desarrollado el sistema AESOP[4], que lleva a cabo una síntesis del habla conversacional (Campbell, 2005) basada en el análisis y la utilización de un enorme corpus oral en japonés que contiene interacciones naturales cotidianas grabadas durante varios años (Campbell, 2004). Sin embargo, el sistema de síntesis expresiva de IBM (Pitrelli et al., 2006) para el inglés americano emplea la voz grabada por un locutor profesional con la intención de generar el estilo correcto adaptándose dinámicamente a la naturaleza del mensaje (buenas o malas noticias, preguntas, disculpa, etc.). Como último ejemplo de esta categoría, debe destacarse *Emospeak*[5], que es la evolución del sistema MARY y que ofrece numerosas mejoras respecto la versión basada en difonemas: incorpora una nueva dimensión emocional (potencia), controla tres niveles de esfuerzo vocal e introduce un nuevo coste en la función de selección de unidades (Schröder et al., 2006).

Modelos ocultos de Markov. Yamagishi et al. (2003) presentan un enfoque alternativo a la síntesis expresiva utilizando síntesis del habla basada en HMM. En esta aproximación, se modelan estadísticamente diferentes estilos del habla y expresiones emocionales y se generan sin necesidad de utilizar reglas heurísticas para controlar la prosodia y otros parámetros del habla sintetizada. Se plantean dos opciones: modelar los estilos de forma individual o de forma simultánea, añadiendo el propio estilo a los otros contextos –fonéticos, prosódicos y lingüísticos– ya existentes (como también proponen Tsuzuki et al. (2004)). Los estilos modelados son el neutro, el alegre, el triste y el agresivo. Posteriormente, Tachibana et al. (2004) proponen un método para generar nuevos estilos a partir de la interpolación de dos estilos ya modelados. Los experimentos realizados mostraron que el habla generada de esta forma transmitía un estilo intermedio a los dos utilizados.

[4] http://feast.atr.jp/AESOP/
[5] http://mary.dfki.de/online-demos/emospeak

Capítulo 4

Corpus oral para la síntesis del habla expresiva

La motivación que llevó al GPMM a la producción de un corpus de habla expresiva en español fue la falta de disponibilidad de un recurso de este tipo que nos permitiese mejorar la naturalidad del sistema de síntesis del habla desarrollado hasta el momento. Como se ha descrito en el apartado 3.1.6, el número de bases de datos existentes en el ámbito nacional era reducido y con unas características que no respondían a nuestras necesidades.

La primera aproximación del autor de la presente tesis a la síntesis del habla expresiva fue a raíz de la colaboración con el LAICOM-UAB, en la cual se utilizó un sintetizador basado en concatenación de difonemas y trifonemas para intentar validar los modelos acústicos que se habían obtenido mediante el análisis del corpus descrito en el apartado 3.1.6.4. Dicho corpus era multilocutor (4 hombres y 4 mujeres), pero de tamaño reducido y, por lo tanto, no apto para su uso en síntesis del habla. El sintetizador de voz utilizado, juntamente con el editor de mensajes orales de voz sintética (EMOVS, Alías y Iriondo (2002)), permitieron generar frases sintetizadas con la prosodia adaptada a los modelos obtenidos para cada emoción. Una de las conclusiones a las que se llegó fue que los parámetros prosódicos utilizados eran insuficientes para simular algunas emociones. Se obtuvieron resultados positivos en la simulación de la tristeza, el enfado y el miedo, pero no se consiguió simular la alegría (Iriondo et al., 2000). La base de datos utilizada por dicho sintetizador solamente constaba de una realización para cada unidad (difonemas y algunos trifonemas), extraída de una grabación de palabras leídas con un estilo neutro.

Dada la situación de los recursos disponibles, se hizo imprescindible abordar la producción de un corpus oral expresivo siguiendo todos los pasos necesarios para garantizar su utilidad en la síntesis del habla expresiva. Este corpus tenía una doble finalidad: el aprendizaje de modelos acústicos por una parte, y su uso como base de datos con las unidades de voz necesarias para el sintetizador por otra.

La construcción de un corpus oral expresivo debe seguir una serie de etapas re-

lacionadas entre ellas. Para el corpus desarrollado en el ámbito de la presente tesis se definieron una serie de etapas que garantizasen, en la medida que fuera posible, una buena calidad de audio, una variabilidad de habla expresiva suficiente que permitiese avanzar en la investigación en síntesis, una organización estructurada que posibilitase su utilización en procesos automáticos y, por último, una validación desde el punto de vista expresivo ya que se trataría de voz grabada por una locutora.

Por lo tanto, el desarrollo del corpus se inició con la determinación de los estilos expresivos y la preparación (diseño) de los textos que se debían grabar (apartado 4.1). En segundo lugar, se llevó a cabo la grabación de los textos diseñados para facilitar la pronunciación de cada estilo por parte de la locutora (apartado 4.2). Una vez realizada la grabación y, antes de seguir con el resto de etapas asociadas a la adecuación del corpus para su uso posterior, se decidió realizar una evaluación subjetiva de una muestra de las locuciones del corpus con el fin de validar su contenido expresivo (apartado 4.3). Después del análisis de los resultados de la prueba subjetiva, se realizó la segmentación en unidades y el correspondiente etiquetado (apartado 4.4), al que prosiguió el análisis acústico de las unidades del corpus (apartado 4.5). La evaluación subjetiva se realizó únicamente sobre una muestra del corpus, ya que una revisión exhaustiva de todo el corpus hubiese sido excesivamente costosa. Se consideró oportuno realizar una validación automática de todo el corpus, que fue la base de la definición de un método para la revisión automática mediante el uso de técnicas de reconocimiento de emociones (apartado 4.6).

La experiencia previa en el diseño y la grabación de un corpus para el catalán (Guaus y Iriondo, 2000) y otro en castellano para la síntesis de previsiones meteorológicas orales (Alías et al., 2005) contribuyó notablemente a la definición y la ejecución de las diferentes tareas necesarias para la producción de este nuevo corpus.

4.1. Diseño del corpus oral expresivo

El primer paso en la producción de un corpus oral consiste en la planificación de las tareas asociadas y el diseño de los elementos que serán la base del material resultante. El diseño del corpus depende de los objetivos que se persigan y de las limitaciones que se puedan aceptar. Los dos apartados siguientes explican, en primer lugar, los objetivos generales relativos a la creación de un corpus oral orientado a la síntesis del habla expresiva y, en segundo lugar, se concretan los pasos seguidos en el diseño del presente corpus.

4.1.1. Objetivos generales

A continuación se detallan los objetivos generales que se deberían alcanzar mediante el desarrollo de un corpus oral expresivo:

1. **Naturalidad / calidad de la grabación**

   La principal característica que debería tener un corpus de habla emocionada es la naturalidad, entendida como la capacidad de transmitir el estado emocional auténtico del hablante (véase el apartado 3.1.2). Como se ha comentado en los apartados 3.1.4 y 3.1.5, el habla espontánea es la más natural, pero su utilización en síntesis del habla presenta diferentes dificultades como la falta de control sobre el contenido y, habitualmente, una calidad de audio insuficiente debido a las condiciones de grabación. Por tanto, un objetivo prioritario es la calidad de la grabación en lo que se refiere a los diferentes aspectos involucrados (equipos de sonido, condiciones acústicas de la sala, profesionalidad del locutor, personal de apoyo, etc.). De todas formas, este objetivo no tiene que ser un impedimento para que las grabaciones realizadas posean un contenido emocional suficiente para conseguir una síntesis del habla expresiva de alta calidad.

2. **Cobertura emocional**

   Otra característica deseable en un corpus de habla emocionada es que presente una amplia representación de emociones, actitudes o estados de ánimos de uno o más hablantes. A priori, podría parecer que es una cuestión que se resolvería mediante la obtención de un corpus de grandes dimensiones recogido en situaciones cotidianas (Campbell, 2002). Sin embargo, una investigación de esta índole puede llegar a requerir varios años únicamente para adquirir el material de voz. Tal volumen de información puede ser muy valioso para conseguir el ambicioso objetivo de aproximar la síntesis de voz al habla conversacional natural (Campbell, 2005). En este sentido, nos marcamos un objetivo menos ambicioso, pero asumible, que fue desarrollar un corpus que nos permitiera dar un paso hacia delante, ya que partíamos únicamente de una experiencia previa en corpus consistentes en la lectura de palabras o frases sin emoción. Se trataba de disponer de un corpus formado por diferentes emociones con un tamaño y cobertura suficientes para poder experimentar en el campo de la síntesis del habla emocional basada en corpus y con un coste económico y un tiempo de desarrollo que el grupo de investigación pudiese asumir.

3. **Cobertura fonética segmental y suprasegmental**

   En la síntesis basada en corpus se requiere una base de datos de habla continua formada por el máximo número de unidades fonéticas y que presente una variación significativa de las características lingüísticas que se deseen reproducir (François y Boëffard, 2002). De igual forma que en el objetivo anterior, será importante controlar el tamaño de la base de datos resultante. Por lo tanto, un objetivo del diseño de un corpus oral será extraer un subconjunto de frases de un amplio corpus textual que den una buena cobertura en cuanto a las unidades fonéticas y a la variabilidad prosódica requerida. En los sistemas de CTH por concatenación, las unidades principalmente utilizadas son los semifonemas, los difonemas y los trifonemas, ya que la concatenación por las partes estacionarias es menos problemática para la mayor parte de los fonemas. La utilización de semifonemas simplifica el problema de la cobertura ya que su número es muy inferior al de difonemas o trifonemas. Sin embargo, su utilización no es adecuada para una síntesis del habla de alta calidad. En el caso de los trifonemas, la cobertura total es prácticamente imposible (Bozkurt et al., 2003). Por lo tanto, el objetivo en el nivel segmental será conseguir una buena cobertura de difonemas y de los trifonemas más necesarios. Un aspecto que debe tenerse en cuenta en lo que se refiere a los requisitos establecidos para conseguir una determinada cobertura fonética es la frecuencia de aparición de los fonemas en una lengua en concreto. Además de garantizar una frecuencia mínima de aparición, se tiene que intentar que las repeticiones de unidades reflejen la frecuencia de aparición propia de esa lengua. En cuanto a la cobertura prosódica, se asegurará una variedad de oraciones enunciativas, interrogativas y exclamativas. Hay que resaltar que se trata de un aspecto muy dependiente de la lengua y que el diseño variará sustancialmente si se pretende desarrollar un sistema multilingüe.

4. **Disponibilidad de corpus textuales adecuados**

   El material textual previo es un elemento clave en el diseño del corpus oral ya que, como se ha comentado en el apartado anterior, debería estar equilibrado segmental y prosódicamente, para ofrecer una buena cobertura. Al considerar emociones, es deseable que los textos de cada subcorpus tengan el contenido semántico adecuado para facilitar la expresión de emociones por parte del locutor. La dificultad del diseño de los textos se multiplica por el número de emociones requeridas (Navas et al., 2006). Por lo tanto, un objetivo importante será conseguir un material textual abundante ya existente del cual se pueda extraer la colección de frases que servirán de base para la grabación del corpus oral. Esta estrategia puede resultar menos costosa que una redacción expresa de los textos.

4.1.2. Enfoque del diseño del corpus oral expresivo

Para la consecución de los objetivos acabados de describir se ha contado con la colaboración de expertos del LAICOM-UAB, quienes nos han asesorado en algunos aspectos del enfoque descrito en los puntos siguientes.

4.1.2.1. Habla emocional estimulada

La consecución de un corpus oral con una amplia variedad expresiva orientado a la investigación en síntesis del habla se ha realizado a partir de la lectura de textos con un contenido semántico adecuado al estilo correspondiente. Para ello, se ha contratado a una locutora profesional capaz de utilizar el estilo expresivo adecuado a cada subconjunto de textos. A fin de mantener la calidad de la señal de voz, todo el corpus se ha grabado en el estudio de grabación de *Enginyeria i Arquitectura La Salle* de la *Universitat Ramon Llull* (EALS-URL). Existe un alto grado de consenso en la comunidad científica en lo que respecta a la utilización de esta estrategia de obtención de habla emocionada para su uso en síntesis (Cowie et al., 2005) aunque, como ya se ha discutido en el apartado 3.1.4, existen también otros enfoques como el de Nick Campbell, que aboga por la construcción de un corpus muy amplio recogido a partir de grabaciones de la vida cotidiana de locutores voluntarios. Por lo tanto, se asume que el entorno controlado de la grabación y el diseño de tales grabaciones pueda limitar el sistema desarrollado, en el sentido de que únicamente llegue a producir estilos de habla formales y pueda también reducir su capacidad de modelar las características de habla espontánea informal (Campbell, 2005).

4.1.2.2. Corpus textual de publicidad

Dada la experiencia del LAICOM-UAB en el uso de la voz en la publicidad, se aprovechó la existencia de un amplio corpus textual de frases publicitarias recopiladas de diarios y revistas que previamente ya estaba organizado en categorías temáticas. Las categorías temáticas escogidas fueron: industria del automóvil, viajes, nuevas tecnologías, educación y cosmética. Según los expertos del LAICOM-UAB, cada una de estas cinco categorías facilitaría la consecución de un estilo de locución propio, lo que permitía la creación de un corpus oral expresivo con una buena cobertura de emociones simuladas. La estrategia que se debía seguir se fundamentaba en el ensayo previo de un determinado estilo a partir del conocimiento experto. Una vez establecidas las características fonéticas de un estilo se procedería a la grabación de los enunciados extraídos del corpus textual publicitario bajo la supervisión de un experto que evitase desviaciones del estilo previamente definido. Por lo tanto, hay que resaltar que la utilización de enunciados provenientes de textos publicitarios tiene como principal objetivo ayudar al locutor a mantener el estilo deseado. En ningún caso se trata de una lectura espontánea de los textos con una libre interpretación por parte del locutor, sino que los estilos se han grabado por bloques y, aunque algún enunciado no sea coherente con el estilo asignado, el locutor debe de ser capaz de mantener la expresividad requerida durante su lectura.

4.1.2.3. Estilos publicitarios

El corpus de textos publicitarios que sirviese de base para la grabación del corpus oral expresivo requería la definición de unos estilos de locución adecuados al contenido. Para ello nos basamos en los estudios de la voz en la publicidad audiovisual (Montoya, 1999, pág. 178), en los que se definió el estilo publicitario como el propio de “aquellas

voces, acusmáticas[1], que interpretan o expresan un texto escrito poniendo énfasis en el ritmo de lectura, en la duración de las pausas antes de nombrar la marca, en la regularidad de los grupos fónicos, en la variabilidad tonal de origen emocional, y en la acentuación de palabras claves. La actitud del locutor, fingida, es básicamente de alegría, o de euforia, de estabilidad emocional o dureza, para provocar un sentimiento determinado en el oyente, logrando así un efecto persuasivo".

Los estilos de locución se establecieron a través de un análisis acústico de diversos anuncios con estereotipos sonoros basados en los rasgos de la personalidad y los estados emocionales. Montoya (2000) definió los estilos siguientes:

1. Estilo de locución alegre: estereotipo de locutor extrovertido/alegre/fascinado

   Este estilo de locución se caracteriza por la poca variación de intensidad, la elevada variación de tono, las pausas cortas y el ritmo rápido. Para marcar los acentos se manipula más el tono que la intensidad, pero van combinados. Utiliza un tono agudo como corresponde al estado de ánimo que quiere transmitir: la alegría.

2. Estilo de locución estable: estereotipo de locutor estable/inteligente/sensitivo/maduro

   Este estilo se caracteriza por un tono grave, que presenta mayor variación en las ramas finales. Se caracteriza también por una regularidad en la duración de las pausas y en la duración de los grupos fónicos. Su ritmo es rápido, aunque con una actitud pausada y tranquila. Los acentos se marcan con una subida y bajada de tono, variando muy poco la intensidad.

3. Estilo de locución duro: estereotipo de locutor dominante/duro

   Este estilo de locución mantiene una intensidad regular, y varia el tono al final de los grupos fónicos. Las pausas son más largas y el ritmo lento. Se caracteriza también por un alargamiento de las consonantes. Además, la voz empleada en este estilo es grave, amenazante, ya que en general mantiene un tono bajo. Para marcar los acentos se realiza una subida de intensidad y de tono, poniendo de relieve con los acentos ciertas sílabas y no otras.

4. Estilo de locución triste (se obtuvo a partir del análisis de una grabación realizada para un doblaje)

   En la voz que se percibe como triste se observan variaciones de intensidad; en cambio, existe poca variación del tono, la velocidad de locución es baja y las pausas son numerosas y más largas.

5. Estilo de locución sensual (modelado a partir del análisis de anuncios de perfumes)

   Para este estilo, la locutora presenta una actitud dulce que se caracteriza por una articulación precisa. Da la sensación que la locutora está haciendo una confidencia o hablando muy cerca de la persona a la que se dirige.

[1] "Acusmático, se dice de una situación de escucha donde, para el oyente, la fuente sonora es invisible; traducción del término *acousmatique* de la Enciclopedia Larousse en línea (`http://www.larousse.fr`)"

4.1.2.4. Obtención de la emoción mediante textos expresivos

Del corpus publicitario se han escogido cinco categorías, a las que se ha asignado un estilo de locución concreto. Esta asignación se sustenta en la existencia de un ensayo previo de cada estilo de locución y en la realización de una supervisión experta durante la grabación, de forma que este tipo de textos simplemente facilita la continuidad del mismo estilo expresivo durante toda la lectura de los enunciados que pertenecen a una misma temática. Las categorías publicitarias y sus estilos asociados son:

1. Nuevas tecnologías: estilo **neutro** (NEU) que transmite una cierta madurez.
2. Educación: estilo **alegre** (ALE) que da sensación de persona extrovertida.
3. Cosmética: estilo **sensual** (SEN) basada en una voz dulce.
4. Automóviles: estilo **agresivo** (AGR) que transmite dureza.
5. Viajes: estilo **triste** (TRI) con cierto aire de melancolía.

De cada categoría se han seleccionado un conjunto de frases mediante un algoritmo de tipo voraz (*greedy* en inglés) (François y Boëffard, 2002) que ha permitido conseguir un equilibrio fonético en cada subcorpus. Este tipo de algoritmos voraces toman decisiones localmente óptimas en cada etapa para aumentar la velocidad pero con la esperanza de encontrar una solución global adecuada. Por lo tanto, la aplicación de este algoritmo al problema planteado conseguirá una solución válida, pero no la óptima (véase el algoritmo 1). Este algoritmo parte de un conjunto de frases C e inicializa el conjunto de salida S al conjunto vacío. De forma iterativa va seleccionando aquella frase de C que maximiza localmente la cobertura deseada y la añade a S. La función *esFactible* pretende evitar frases con excepciones o muy similares a las ya incorporadas a S. La función *esSolucionFinal* comprueba si el subconjunto S cumple los requisitos iniciales R (p.ej. la cobertura fonética establecida).

Con el fin de favorecer la generación del corpus y su mejor aprovechamiento, además de buscar un equilibrio fonético, se han incorporado a la función *esFactible* los criterios siguientes:

- Evitar frases que contengan excepciones (palabras extranjeras, abreviaturas) que dificulten el proceso de transcripción fonética y etiquetado automático.
- Penalizar la aparición de frases similares a las previamente seleccionadas debido a la aparición de alguna unidad nueva.
- Penalizar la selección de frases excesivamente cortas o largas para facilitar la interpretación por parte del locutor. La activación de este criterio únicamente permite la selección de frases cuya longitud en número de fonemas se encuentre entre dos umbrales que se pueden configurar manualmente.

Algoritmo 1 Algoritmo *greedy* para la selección de frases

```
AlgoritmoGreedy(C: Conjunto Frases, R: Requisitos)
S := ∅;
while C ≠ ∅ and ¬solucionEncontrada do
  f := seleccionaFraseCandidata(C);
  if esFactible(f ∪ S) then
    S := S ∪ f;
    C := C − f;
    if esSolucionFinal(S, R) then
      solucionEncontrada = TRUE;
    end if
  end if
end while
return S; {devuelve la solución}
```

Una de las entradas al sistema de selección de frases es el inventario de segmentos (fonemas y alófonos) que pueden aparecer en la transcripción fonética de las frases. La transcripción fonética de todas las frases del conjunto inicial C se realiza mediante la herramienta desarrollada por el GPMM (véase la descripción en el apartado E.1) que utiliza el inventario de segmentos descrito en el apartado B.1.

Para optimizar el proceso de selección, los segmentos requeridos se han ordenado de menor a mayor frecuencia de aparición, con la finalidad de que el algoritmo de *greedy* comience a elegir frases que contengan fonemas con menos probabilidad de aparecer. Para conocer a priori la frecuencia de aparición de los fonemas en español se ha consultado el estudio comparativo de Pérez (2003). Se puede consultar una muestra de las frases escogidas en el apartado B.1 del anexo de la presente memoria.

La distribución fonética por subcorpus y el total de las vocales se muestra en la figura 4.1, en las que se distingue entre vocales átonas (/a/ /e/ /i/ /o/ /u/), tónicas (/A/ /E/ /I/ /O/ /U/) y semivocales (/j/ /w/). El porcentaje de consonantes se muestra en la figura 4.2.

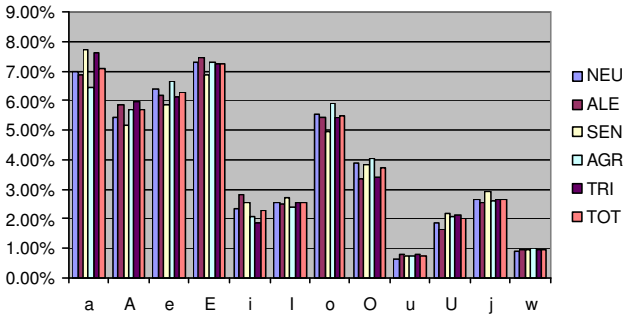


Figura 4.1: Distribución de las vocales por estilo y para los cinco estilos (TOT)

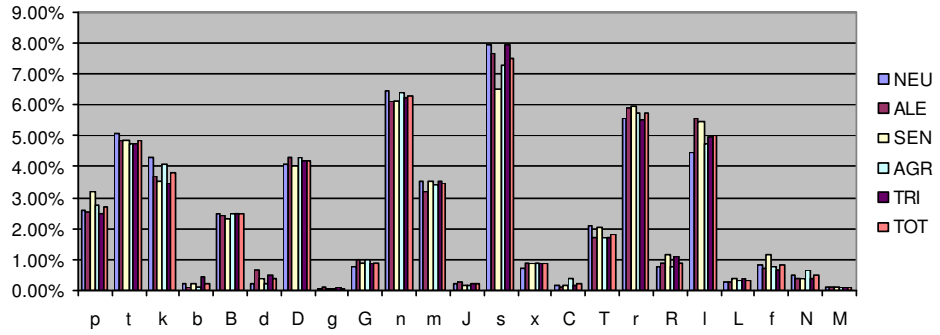


Figura 4.2: Distribución de las consonantes por estilo y para los cinco estilos (TOT)

Los resultados de frecuencia de aparición de segmentos correspondientes a los textos seleccionados para la grabación del corpus son muy similares al promedio de los cinco estudios presentado en Pérez (2003). En la tabla 4.1, se compara la frecuencia de los fonemas vocálicos. Los valores correspondientes al total del corpus diseñado incluyen para cada vocal la suma de las unidades átonas y tónicas. La tabla 4.2 muestra los resultados para los fonemas consonánticos. En la presente tabla, los valores para los fonemas /b/, /d/ y /g/ representan la suma de las frecuencias de ambos alófonos de cada fonema, el oclusivo y el aproximante, ya que en el estudio de referencia no se diferenciaban.

Tabla 4.1: Comparación de la frecuencia de aparición de las vocales en el total del corpus diseñado y el promedio de los cinco estudios presentado en Pérez (2003)

	/a/	/e/	/i/	/o/	/u/
Corpus diseñado	12,74	13,56	6,13	9,24	2,74
Pérez (2003)	13,27	13,13	6,32	9,71	2,32

Tabla 4.2: Comparación de la frecuencia de aparición de las consonantes en el total del corpus diseñado y el promedio de los cinco estudios presentado en Pérez (2003)

	/p/	/t/	/k/	/b/	/d/	/g/	/n/	/m/	/J/
Corpus diseñado	2,70	4,82	3,84	2,67	4,59	0,99	6,27	3,44	0,22
Pérez (2003)	2,66	4,66	4,02	2,66	4,58	1,02	5,3	2,73	0,28
	/s/	/x/	/C/	/T/	/r/	/R/	/l/	/L/	/f/
Corpus diseñado	7,51	0,85	0,24	1,83	5,72	0,92	4,99	0,32	0,81
Pérez (2003)	8,72	0,65	0,34	1,89	4,48	0,69	4,86	0,57	0,74

4.1.2.5. Lista de palabras portadoras

Con la finalidad de garantizar la aparición de todos los difonemas y los trifonemas utilizados se creó una lista de palabras portadoras que los contenían. Cada unidad de palabras portadoras está formada bien por una sola palabra si en su interior contiene la unidad requerida, bien por dos palabras si la unidad aparece por contacto del final de la primera palabra con el inicio de la segunda. El objetivo de esta lista de palabras portadoras es garantizar a priori la presencia de todas las unidades en cada subcorpus. Además, permite realizar comparaciones directas entre los parámetros acústicos de los 5 estilos, aunque solamente sea a nivel segmental. La lista actual, definida a partir de la revisión de una lista anterior que constaba de 698 palabras, contiene 1.250 palabras. La nueva tabla distingue entre vocales tónicas y átonas y, además, se han añadido algunas unidades que no estaban definidas. En el apartado B.3 se puede consultar la lista completa de palabras con su transcripción fonética y la unidad que contiene.

4.2. Grabación

Tal y como se ha justificado en el apartado 4.1, la obtención del habla expresiva se ha realizado mediante la lectura por parte de una locutora de los textos definidos a tal efecto. Un requisito del corpus oral es que disponga de una calidad de audio excelente para su posterior uso en un sistema de síntesis del habla.

4.2.1. Instalaciones y equipo de grabación

La grabación del corpus oral se ha llevado a cabo en las instalaciones del Departamento de Tecnologías Audiovisuales de EALS-URL, concretamente en el estudio de grabación. Dicho estudio consta de dos salas: la sala de control, que dispone del equipo necesario para la mezcla y producción de lo que se graba, y la sala de grabación. Ambas salas están tratadas acústicamente para ofrecer una respuesta adecuada. Concretamente, la sala de grabación tiene forma irregular con una planta de 5 por 4 metros y una altura de 3,5 metros. El tiempo de respuesta de la sala es de unos 0,8 segundos, pero la situación relativa entre el locutor y el micrófono garantiza la ausencia de ecos audibles.

Se ha utilizado un micrófono de condensador (*AKG C-414*) con una respuesta prácticamente plana (2 dB en el rango de 20-20000 Hz) y una relación señal a ruido de 80 dBA SPL. La grabación se ha realizado directamente en un disco duro mediante la plataforma digital *Pro Tools 5.1* instalada en un ordenador *Mac G5* que utilizaba una consola digital *Yamaha 02R*. La digitalización de la señal se ha llevado a cabo con una frecuencia de muestreo de 48 KHz y una cuantificación de 24 bits en ficheros del tipo WAV.

4.2.2. Dinámica de las sesiones de grabación

Las sesiones de grabación han seguido un protocolo preestablecido con el fin de minimizar errores que puedan causar deficiencias en procesos posteriores, como en la segmentación y el etiquetado del corpus o en la síntesis del habla. Este protocolo requiere un equipo de personas formado por un técnico de sonido, un experto en comunicación audiovisual, un técnico de control y el locutor o locutora.

El técnico de sonido es el responsable de ajustar la plataforma de grabación y la posición del micrófono de forma óptima para una grabación de voz. Es importante que al iniciar una nueva sesión en un día diferente se mantengan de la forma más similar posible las condiciones de la sesión anterior.

El experto en comunicación audiovisual ensaya con el locutor los diferentes estilos y lo corrige en el caso de que se desvíe del modelo deseado (véase el apartado 4.1.2.3).

El técnico de control tiene la misma lista de frases que el locutor y verifica la coincidencia entre el texto y la locución. En función del tipo de incidencia, simplemente realiza una anotación en la frase correspondiente de la lista o requiere su repetición.

4.3. Evaluación subjetiva

La evaluación subjetiva es una herramienta que nos permite validar un corpus de habla emocional grabado por un actor o locutor profesional. El objetivo de esta evaluación consiste en validar, tomando como referencia las opiniones de una muestra de oyentes, el contenido emocional o expresivo simulado en la grabación. La validación de la expresividad del corpus se completará con unas pruebas de identificación automática sobre el corpus entero (evaluación objetiva).

4.3.1. Diseño del test

No se ha planteado una evaluación exhaustiva del corpus, ya que, dada su extensión, la evaluación completa sería un proceso excesivamente largo (el tamaño del corpus es de un total de 4.638 frases). Por este motivo, para cada estilo se han escogido aleatoriamente 96 frases, lo que representa un total de 96*5 = 480 frases. Estas 480 frases se han dividido en 4 subconjuntos de test, cada uno de los cuales consta de 120 frases. A cada evaluador se le ha asignado un par ordenado de estos 4 subconjuntos, lo que da lugar a 12 pruebas diferentes. La idea de asignar pares ordenados pretende compensar el hecho de que el segundo subconjunto pueda ser más fácil de evaluar debido a la experiencia adquirida por el evaluador (p.ej. habrá usuarios que evaluarán primero las frases correspondientes al test 1 y después al 3, y otros que primero realizarán el 3 y después el 1).

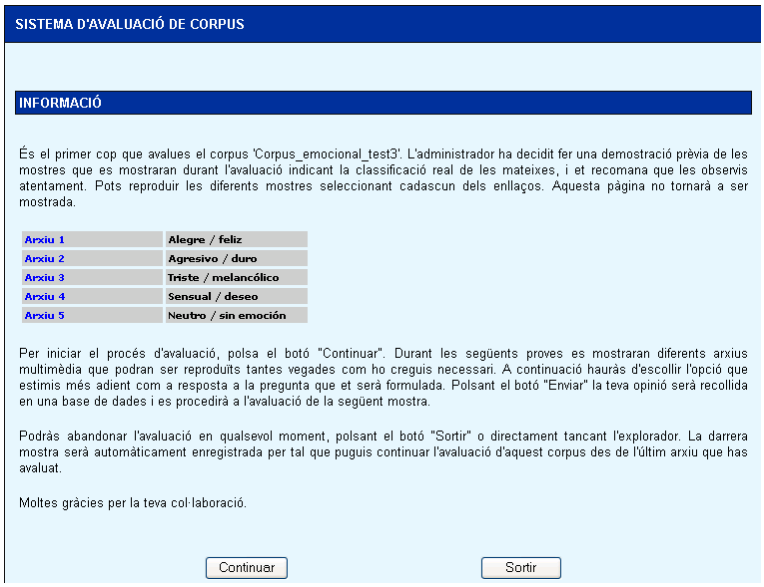
Se ha diseñado una evaluación de respuesta forzada a la pregunta: *"¿Qué estado emocional te transmite la voz de la locutora en esta frase?"*. Las posibles respuestas son los 5 estilos del corpus más una opción *"No lo sé / Otro"*, que se añade con el objetivo de no forzar una respuesta insegura o errónea en aquellos casos difíciles de identificar, aunque, como se indica en Navas et al. (2006), se corra el riesgo de que algunos evaluadores abusen de esta respuesta para acelerar la conclusión del test.

4.3.2. Proceso de evaluación

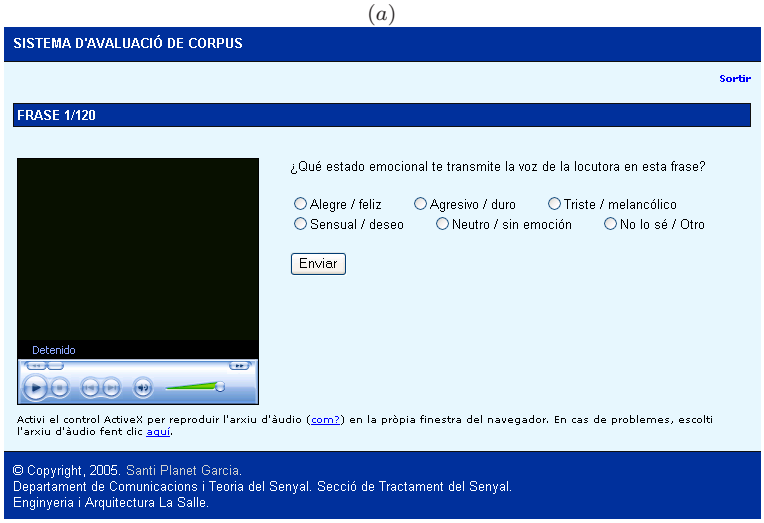
El proceso de evaluación se ha llevado a cabo mediante una plataforma web desarrollada por Santiago Planet del GPMM para realizar este tipo de pruebas. Se trata de una herramienta que permite configurar la página inicial para proporcionar las instrucciones del proceso evaluador y, si es necesario, incluir algunas muestras de ejemplo antes de iniciar el test (véase la figura 4.3a). Una vez iniciado el test, se suceden las páginas de evaluación (véase en ejemplo en la figura 4.3b), en las que se permite al usuario escuchar las frases tantas veces como sea necesario y marcar la opción escogida. También permite cerrar la sesión antes de finalizarla y reanudarla en otro momento. Los resultados se guardan automáticamente en una base de datos para su posterior análisis.

Los evaluadores han sido, en su gran mayoría, estudiantes o profesores vinculados a EALS-URL. La solicitud de voluntarios se ha realizado mediante un envío de correo electrónico a 240 personas, cada una de las cuales tenía asignada una de las 12 pruebas en

las que se dividió la evaluación (apartado 4.3.1). El número final de evaluadores que han colaborado es el siguiente: 25 han completado la prueba asignada (240 frases) y 13 han finalizado únicamente la mitad de la prueba (120 frases). Inicialmente, se han estudiado los resultados de los evaluadores que han completado la prueba asignada, para poder comparar la influencia del orden en que se realizan los dos tests.



(a)



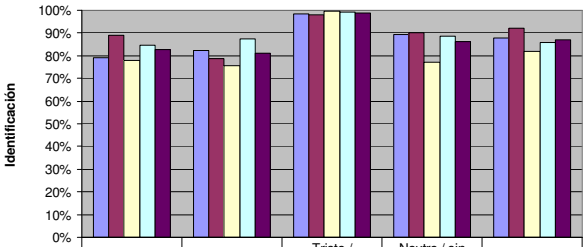
(b)

Figura 4.3: Pantalla inicial de la plataforma de test (a). Pantalla de respuesta forzada de la plataforma de test para un ejemplo concreto (b)

4.3.3. Resultados

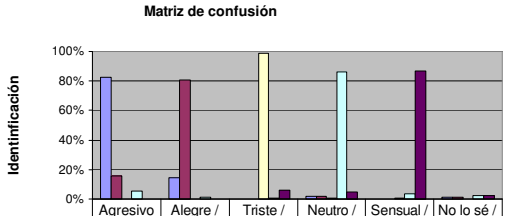
Los resultados de la evaluación subjetiva muestran que todos los estilos se han identificado en un alto porcentaje. La figura 4.4 muestra el porcentaje de identificación por estilo y test, siendo el estilo triste el que obtiene claramente una identificación superior (98.8 % de media), seguido por los estilos sensual (86.8 %) y neutro (86.4 %) y, finalmente, los estilos agresivo (82.7 %) y alegre (81 %). La identificación mayoritaria del estilo triste es la más habitual en los estudios de percepción del habla emocionada, debido a que se diferencia claramente de los otros estilos por su tono medio más bajo, la escasa variabililidad del tono y la ralentización del ritmo del habla.

La matriz de confusión (véase la figura 4.5) indica que los estilos que se han confundido mayoritariamente son el agresivo con el alegre (14.2 % de las frases del subcorpus agresivo identificadas como alegre) y viceversa (15.6 % de las frases del subcorpus alegre identificadas como agresivo). Además, se observa que los estilos neutro y sensual siguen un patrón parecido: *i)* se confunden ligeramente entre ellos y *ii)* cada uno se confunde ($> 5\,\%$) con otro estilo (neutro por agresivo y sensual por triste).



	Agresivo /duro	Alegre / feliz	Triste / melancólico	Neutro / sin emoción	Sensual / deseo
Test1	79.2%	82.2%	98.6%	89.4%	87.8%
Test2	88.9%	78.6%	98.1%	90.3%	91.9%
Test3	78.1%	75.6%	99.4%	77.2%	81.9%
Test4	84.5%	87.5%	99.1%	88.7%	85.7%
Media	82.7%	81.0%	98.8%	86.4%	86.8%

Figura 4.4: Porcentaje de identificación en los 4 tests y promedio total de los 25 evaluadores



	Agresivo /duro	Alegre / feliz	Triste / melancóli	Neutro / sin	Sensual / deseo	No lo sé / Otro
Agresivo /duro	82.7%	14.2%	0.1%	1.8%	0.1%	1.1%
Alegre / feliz	15.6%	81.0%	0.1%	1.9%	0.2%	1.2%
Triste / melancólico	0.0%	0.0%	98.8%	0.5%	0.6%	0.1%
Neutro / sin emoción	5.3%	1.3%	0.7%	86.4%	3.6%	2.7%
Sensual / deseo	0.0%	0.1%	5.7%	4.7%	86.8%	2.6%

Figura 4.5: Histograma y matriz de confusión de los resultados promediados de los 4 tests de identificación. Las columnas indican el estilo identificado por los usuarios.

También se ha estudiado la influencia del orden en la realización del test. En general, la segunda ronda de test obtiene mejores resultados, especialmente en los estilos agresivo, sensual y neutro (véase el diagrama de caja[2] de la figura 4.6). Este fenómeno se debe seguramente al entrenamiento previo que le ha supuesto el primer test para el evaluador. Un resultado contrario se hubiese podido deber al efecto de la fatiga en el evaluador. Seguramente el diseño de la interfaz, que permite abandonar la prueba y reanudarla posteriormente, ha mitigado este efecto. El tiempo aproximado de realización del test ha sido de unos 20 minutos para cada ronda.

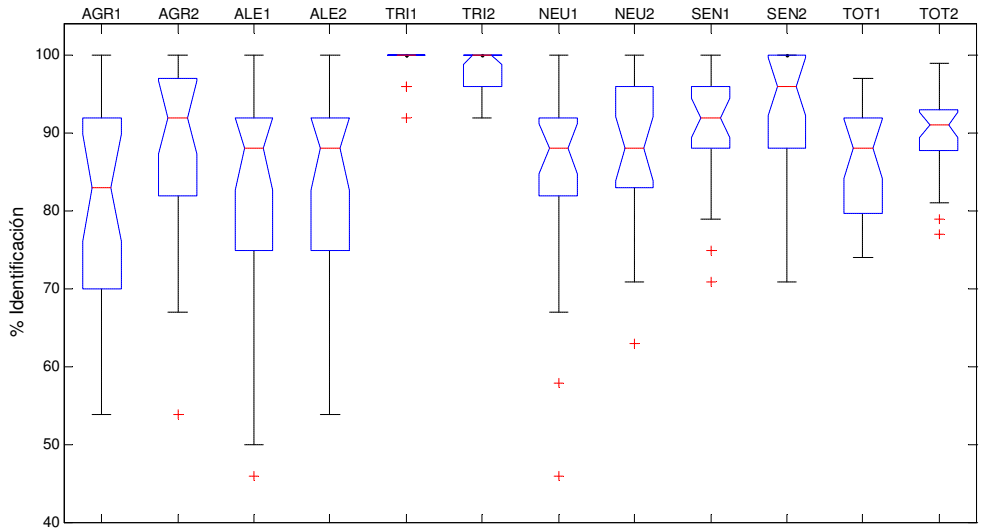


Figura 4.6: Diagrama de caja comparativo de los porcentajes de identificación de cada estilo agrupados de dos en dos según correspondan a resultados del primer test (AGR1, ALE1, etc.) o del segundo (AGR2, ALE2, etc.). El último par corresponde al promedio acumulado de todos los estilos.

[2]Las cajas presentan líneas en el cuarto inferior (mediana de la mitad más pequeña), la mediana y el cuarto superior (mediana de la mitad más grande). El ancho de caja (cuarto superior menos cuarto inferior) es una medida de la dispersión de los datos. Los bigotes son líneas que se extienden desde cada final de caja hasta los datos (superior e inferior) más alejados de la mediana y que no se consideran atípicos. Las observaciones más allá de 1,5 veces el ancho de caja del cuarto más cercano son valores atípicos (Devore, 2005).

4.4. Segmentación y etiquetado

El corpus se ha estructurado en frases y palabras aisladas y, por lo tanto, se debe procesar la grabación original para obtener únicamente la mejor versión de cada una en caso de repetición por parte de la locutora. Para este proceso se ha utilizado la herramienta de reconocimiento del habla *Hidden Markov Model Toolkit (HTK)*[3], configurada con una gramática basada en una transcripción fonética automática del texto proporcionado a la locutora en el momento de la grabación. De esta forma, un bloque de texto grabado se puede segmentar en frases mediante un alineamiento forzado. Una vez segmentado automáticamente el archivo máster, es necesario revisar manualmente el resultado y corregir algunas frases que contienen errores de segmentación. Los errores se han debido principalmente a la falta de coherencia entre el pausado realizado por la locutora y los signos de puntuación. En los casos en que aparecen silencios en el fichero de audio sin el correspondiente signo de puntuación, se han modificado el texto y la transcripción fonética. De forma inversa, aquellas pausas no realizadas por la locutora y que, en cambio, estaban marcadas por un signo de puntuación, se han solucionado quitando el signo de puntuación del texto y de la transcripción fonética. En esta tarea manual han colaborado diferentes miembros del GPMM.

La tabla 4.3 muestra un resumen del número de frases[4] y de palabras portadoras que componen cada estilo, junto con el tiempo total de voz grabada del corpus. Las frases propias son diferentes para cada estilo y su contenido semántico está relacionado con los estilos publicitarios según se ha descrito en el apartado 4.1.2.3. Las palabras portadoras incluyen la unidad fonética que se desea obtener mediante su grabación (véase el apartado 4.1.2.5). Aunque las 1250 palabras portadoras son las mismas para los cinco estilos, su pronunciación intenta reflejar el estilo correspondiente y, por tanto, la duración es ligeramente diferente.

Tabla 4.3: Resumen del contenido del corpus una vez segmentado en frases y palabras portadoras.

Estilo	Frases propias (núm. / tiempo)	Palabras portadoras (núm. / tiempo)
Neutro	833 / 50 min	1250 / 22 min
Alegre	916 / 56 min	1250 / 25 min
Sensual	841 / 51 min	1250 / 31 min
Triste	1000 / 86 min	1250 / 25 min
Agresivo	1048 / 84 min	1250 / 24 min

El análisis acústico del corpus de voz se basa en el etiquetado fonético, que consiste en una secuencia de marcas que delimitan el inicio y final de los segmentos. La segmentación se ha llevado a cabo mediante un alineamiento forzado con HMM utilizando también la herramienta HTK. Esta segmentación se utilizará en el siguiente paso (análisis acústico) cuando sea necesario disponer de parámetros acústicos segmentales. En primer lugar se

[3] http://htk.eng.cam.ac.uk/

[4] Actualmente los estilos neutros, alegre y sensual se han ampliado para equiparar la duración de la muestra a la de los estilos agresivo y triste. En el momento de la redacción de esta tesis, no se disponía todavía de la segmentación y el etiquetado de esta ampliación del corpus.

lleva a cabo un entrenamiento de los HMM únicamente a partir de los archivos de voz y sus respectivas transcripciones fonéticas. Entonces, se dispone de una primera segmentación, parte de la cual se revisa manualmente y se ajustan la marcas de segmentación con la ayuda de una herramienta gráfica. Con los enunciados revisados, se repite el entrenamiento pero esta vez también se proporcionan las marcas de segmentación.

4.5. Análisis acústico

Las características prosódicas (F_0, energía, duración segmental y de las pausas) están relacionadas con el habla emocionada (Cowie et al., 2001). A continuación se explica el análisis acústico automático llevado a cabo a partir de la información previa obtenida mediante la segmentación y el etiquetado fonético del corpus.

4.5.1. Parámetros de frecuencia fundamental

El análisis de parámetros de F_0 se ha realizado sobre el resultado del marcador de F_0 descrito en Alías et al. (2006). Este marcador tiene la particularidad de que en las zonas carentes de sonoridad y en los silencios asigna marcas interpoladas respecto las zonas sonoras vecinas, mejorando así las marcas en las zonas en las que la señal no presenta una periodicidad clara. Para cada frase se obtienen tres vectores de valores locales de F_0 (un vector completo, otro que excluye los silencios y los sonidos sordos, y un tercero únicamente para las vocales tónicas). En el presente trabajo se ha utilizado el etiquetado del corpus para generar el vector que excluye los silencios y los segmentos sordos, así como para generar el vector que incluye valores únicamente de las vocales. En el caso de utilizar habla sin segmentar fonéticamente se requeriría de un detector de actividad de voz (VAD) y un detector de sonoridad (V/UV) como en Navas et al. (2006). Para el uso de la información de las vocales tónicas se requeriría de un detector de acento. Además, para los valores de F_0 se ha utilizado una representación lineal y logarítmica.

4.5.2. Parámetros de energía

Las locuciones se han analizado con ventanas de 20 ms cada 10 ms calculando la energía media para cada trama. Se calcula la energía en unidades *rms* y en decibelios (dB). Siguiendo la misma idea que para F_0, se han generado tres vectores (completo, excluyendo silencios y únicamente con datos de las vocales tónicas).

4.5.3. Parámetros relacionados con el ritmo

La duración de los sonidos es un aspecto importante en la expresión oral de emociones. Algunos estudios omiten este parámetro por la dificultad de obtenerlo automáticamente (Navas et al., 2006). En el presente trabajo hemos incorporado la duración segmental (gracias al etiquetado del corpus) para disponer de conjuntos de datos con y sin esta información y poder contrastar su relevancia.

El modelado de la duración en sistemas de CTH se ha basado habitualmente en la medida *z-score*, ya utilizada por Campbell (1990), para predecir la duración individual de los segmentos y controlar su alargamiento o reducción con el fin de modificar la velocidad del habla. Como en Schweitzer y Möbius (2003), se utiliza el *z-score* para el análisis de la

estructura temporal del habla:

$$z\_score = \frac{dur(ms) - \mu}{\sigma} \tag{4.1}$$

donde μ y σ son la media y la desviación estándar respectivamente del segmento correspondiente, estimadas del corpus entero. Por lo tanto, uno de los elementos que configuran el ritmo de una frase se representa por un vector con el *z-score* de cada segmento. Además, se genera otra versión de este vector únicamente con los valores en las vocales tónicas.

Finalmente, se calculan dos parámetros relacionados con el pausado para cada frase. Estos parámetros son el número de pausas por unidad de tiempo y el porcentaje de tiempo de silencio respecto a la duración total del enunciado. El objetivo de estos parámetros es representar la frecuencia y la duración de las pausas.

4.6. Validación objetiva de la expresividad del corpus

El objetivo de los experimentos descritos en el presente apartado consiste en validar el contenido expresivo del corpus mediante técnicas de identificación automática de emociones utilizando diferentes técnicas de minería de datos aplicadas sobre medidas estadísticas de los parámetros acústicos de las frases. Los motivos que han llevado a la necesidad de validar a posteriori el contenido expresivo de las frases son los siguientes:

- La utilización de una locutora profesional para la producción de voz emocionada tiene el inconveniente de que ciertas frases puedan carecer de la expresividad necesaria. Por lo tanto, aquellas frases con un contenido expresivo diferente al deseado no serán útiles para los diferentes usos en la síntesis del habla.
- Como se ha comentado en el apartado 4.3, una revisión exhaustiva de todo el contenido del corpus sería muy costosa y, por lo tanto, sería de mucho interés desarrollar un sistema automático para llevar a cabo esta tarea. Además, este desarrollo aporta la posibilidad de utilizar la misma metodología en la creación de nuevos corpus orales.
- Dado que el objeto del presente corpus es la síntesis del habla, este se utilizará en un estudio centrado en el oyente, donde lo importante es la capacidad de simular emociones a través del habla. Por consiguiente, la percepción subjetiva será muy importante, y debe, por ello, guiar el proceso de validación automático.

Se han llevado a cabo tres experimentos que han ido evolucionando progresivamente hasta conseguir un sistema capaz de generar automáticamente una lista de frases con un contenido expresivo diferente al deseado.

4.6.1. Evaluación objetiva preliminar

El objetivo del primer experimento de validación automática ha consistido en aplicar técnicas de identificación automática de emociones utilizando diferentes algoritmos de minería de datos aplicadas sobre un conjunto de medidas estadísticas de los parámetros prosódicos (véase el apartado 4.5) en el nivel de la frase.

4.6.1.1. Características y conjunto de datos

La información prosódica de una frase se ha representado por las secuencias de valores de F_0 (lineal y logarítmica), de energía (lineal y dB) y de las duraciones normalizadas (*z-score*) de cada segmento. Como se ha explicado en el apartado 4.5, para cada frase se calculan tres secuencias de F_0, tres de energía y dos de duración. Cada secuencia se repite con las diferentes unidades de medida de su parámetro. Para cada secuencia, además, se calcula la primera y segunda derivada discreta, teniendo en cuenta que cada secuencia

tiene un valor numérico por segmento. Para todas estas secuencias numéricas se calculan los siguientes datos estadísticos: la media, la varianza, el valor máximo, el valor mínimo, el rango, el sesgo, la curtosis, los tres cuartiles y el rango intercuartílico. Considerando también los dos parámetros del pausado (véase el apartado 4.5.3), hacen un total de 464 parámetros por frase (véase la tabla 4.4). A modo de ejemplo, para la fila de F_0 se observa que el resultado es de 198 parámetros por frase, resultado del producto de 2 tipos de unidad (lineal y logarítmico), 3 secuencias (completa, sin silencios ni consonantes sordas, sólo vocales), 3 funciones (la secuencia y la primera y segunda derivadas discretas) y 11 medidas estadísticas.

Tabla 4.4: Desglose de los parámetros usados en la representación prosódica de cada locución para el conjunto de datos de partida (Data1)

	Unidades	Secuencias	Funciones	Medidas estadísticas	Total por frase
F_0	2	3	3	11	198
Energía	2	3	3	11	198
Duración	1	2	3	11	66
Pausado	2	-	-	-	2
TOTAL					464

Este conjunto inicial de datos, denominado Data1, se ha dividido en diferentes subconjuntos siguiendo diferentes estrategias para estudiar la posible reducción del número de parámetros, seleccionando aquellos que son más significativos desde el punto de vista expresivo. El diagrama de la figura 4.7 muestra los conjuntos de datos que se han generado a partir de Data1, indicándose el tipo de reducción efectuada. Un primer criterio para reducirlo es prescindir de la segunda derivada de Data1, obteniéndose así Data2. En segundo lugar, los experimentos preliminares han mostrado que con el uso de las versiones logarítmicas de la F_0 y la energía se consiguen mejores resultados. Por esta razón, se han generado dos nuevos conjuntos de datos con las versiones logarítmicas de F_0 y de la energía. Cada uno de estos conjuntos (Data1L y Data2L) se ha dividido en dos nuevos conjuntos considerando únicamente las secuencias que contienen todos los fonemas y alófonos (Data1LC y Data2LC) o únicamente las secuencias con las vocales tónicas (Data1LS y Data2LS). Estos dos últimos conjuntos se han generado para estudiar si la información contenida en las vocales tónicas es suficiente para distinguir los diferentes estilos.

Además, se ha realizado una reducción automática de los dos conjuntos iniciales (con y sin segunda derivada discreta) por medio de la combinación de un evaluador de atributos y un método de búsqueda implementados por Weka[5] (Witten y Frank, 2005), obteniéndose de este modo Data1G y Data2G. El evaluador de atributos toma un subconjunto de atributos y retorna una medida numérica que guía la búsqueda. Se ha escogido la función *CfsSubsetEval*, que valora simultáneamente la habilidad predictiva de cada atributo del conjunto de forma individual y el grado de redundancia entre ellos, prefiriendo conjuntos de atributos altamente correlacionados con la clase, pero con baja intercorrelación. Como algoritmo de búsqueda se ha escogido la función *GeneticSearch*, que utiliza un algoritmo genético simple basado en Goldberg (1989). Se han utilizado los valores por defecto de la función: el tamaño de la población (20), el número máximo de generaciones

[5] http://www.cs.waikato.ac.nz/ml/weka/

(20) y las probabilidades de cruce (0,6) y de mutación (0,33). Esta reducción es independiente del algoritmo de clasificación utilizado posteriormente y, por lo tanto, en este primer experimento, todos los métodos de clasificación se han probado con los mismos conjuntos de datos.

Finalmente, se han generado dos conjuntos de datos similares a los presentados en Navas et al. (2006) con la finalidad de estudiar las consecuencias de omitir los parámetros de ritmo. Data1N se ha generado a partir del cálculo de 7 medidas estadísticas (valores medio, máximo y mínimo, desviación estándar, rango, sesgo y curtosis) de las secuencias de F_0 y energía, ambas en versión lineal y logarítmica, y de las derivadas primera y segunda. La F_0 se ha calculado solamente en los segmentos sonoros, y para la energía se han excluido los silencios. Data1NG se ha generado aplicando la misma técnica de selección de atributos que a los conjuntos Data1 y Data2, obteniéndose un subconjunto de 39 atributos, que es el mismo número que el obtenido en Navas et al. (2006), aunque con un método diferente. Se desconoce la coincidencia cualitativa de ambos conjuntos de datos y, por consiguiente, los resultados obtenidos para este conjunto de datos sólo son orientativos y no pretenden comparar ambos sistemas, sino únicamente valorar el funcionamiento con atributos relativos a la F_0 y la energía, que son más fáciles de extraer automáticamente desde la señal de voz que los atributos relacionados con el ritmo.

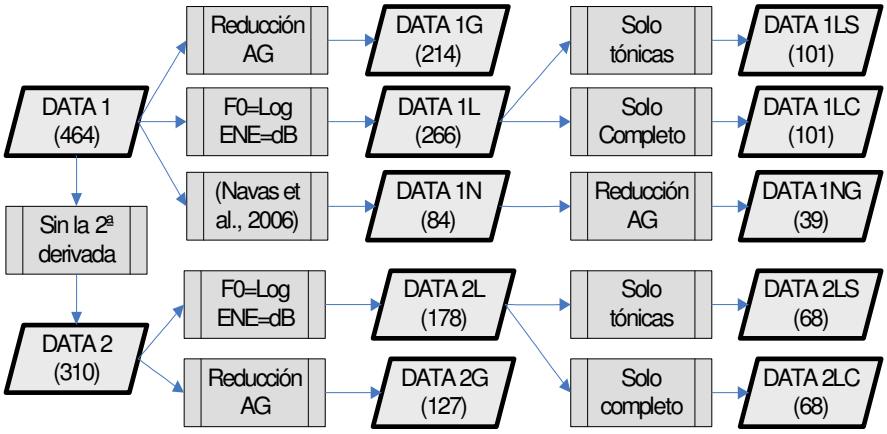


Figura 4.7: Generación de diferentes conjuntos de datos

4.6.1.2. Algoritmos de clasificación

Existen numerosos esquemas de aprendizaje automático que pueden utilizarse en la tarea de clasificar el estilo o la emoción de un enunciado a partir del análisis de la prosodia del habla. Se ha utilizado como base un experimento a gran escala de reconocimiento de emociones en el que se combinaron un gran espacio de parámetros con un gran número de algoritmos de aprendizaje automático (Oudeyer, 2003).

Se han utilizado los algoritmos de clasificación supervisada descritos a continuación,

tal y como se implementan en el *software* Weka[6], siguiendo una estrategia de validación cruzada en 10 bloques. Es decir, el conjunto de datos se divide aleatoriamente en 10 bloques; se destinan 9 al entrenamiento y el restante a test. El proceso se repite un total de 10 veces de forma que todos los elementos del conjunto de datos forman parte del conjunto de test en una de las 10 iteraciones. Los algoritmos utilizados son:

- ***J48*** implementa la versión pública del algoritmo de clasificación basada en árboles de decisión C4.5 revisión 8, previa a la comercialización de la versión C5.0 (Quinlan, 1993, citado por Witten y Frank, 2005, p. 198). Estos árboles clasifican un nuevo caso mediante la evaluación, en cada nodo del modelo, de los parámetros que definen el caso que se pretende clasificar. Los casos que, partiendo de la raíz, llegan a una determinada hoja reciben la clasificación que la hoja indica.

- ***DecisionTable (DT)*** implementa un clasificador basado en el resultado mayoritario de una tabla de decisión (Kohavi, 1995, citado por Witten y Frank, 2005, p. 408). Los datos se representan mediante dos componentes: un esquema, que es un conjunto de atributos, y un cuerpo formado por casos etiquetados con los atributos que contiene el esquema. Dado un caso no etiquetado, se buscan las coincidencias exactas en la tabla utilizando únicamente los atributos del esquema. Si no hay ninguna coincidencia, se retorna la clase mayoritaria de la tabla; en el caso de encontrar instancias idénticas, se retorna la clase con mayor número de coincidencias. Para decidir qué atributos formarán el esquema, se ha utilizado el método *leave-one-out*[7].

- ***PART*** —de *Partial Decision Trees*— es un algoritmo de creación de reglas a partir de árboles de decisión que siguen la heurística de C4.5. El algoritmo construye una regla, elimina las instancias que cubre dicha regla, y continua creando reglas recursivamente para el resto de instancias hasta que quedan todas cubiertas. La creación de una regla individual se basa en la creación de un árbol de decisión con poda aplicado al conjunto de instancias activo. La hoja que cubre más casos se convierte en regla, y el resto del árbol se descarta (Frank y Witten, 1998).

- ***Ib1, Ibk*** son clasificadores basados en ejemplos (*Instance Based*), que almacenan las muestras etiquetadas de entrenamiento directamente. Para clasificar una nueva muestra se emplea una función de distancia para evaluar qué muestra o muestras del conjunto de entrenamiento son las más próximas a ella. Para el algoritmo IB1, la nueva muestra se clasifica con la etiqueta de la muestra más cercana. En el caso del algoritmo IBk se observan las clases de los k vecinos más próximos y la clasificación final se decide según la votación mayoritaria (Witten y Frank, 2005). El algoritmo *IBk* presenta una característica adicional: la elección del número óptimo k mediante validación cruzada (concretamente, *leave-one-out*). A pesar de ser costoso desde el punto de vista computacional, mejora notablemente los resultados de IB1.

- ***Naive Bayes (NB)*** (John y Langley, 1995) es un clasificador probabilístico que parte de la premisa de que cada par parámetro-valor de un mismo ejemplo es independiente del resto. A cada par parámetro-valor se le asigna una probabilidad de

[6] Las funciones de Weka se han ejecutado mediante llamadas desde Matlab® que utilizan tecnología Java.
[7] La validación cruzada *leave-one-out* consiste en eliminar cada instancia y entrenar con el resto.

pertenencia a una clase. Para ello se divide el número de ejemplos de cada clase en los que aparece ese par entre el número de ejemplos que pertenecen a esa clase. Para clasificar un caso nuevo se calcula la probabilidad de pertenencia de ese caso a cada clase, clasificándolo en la clase donde dicha probabilidad sea mayor, adoptando pues un criterio de estimación máxima a posteriori. Esta probabilidad de pertenencia se calcula como el producto de la probabilidad de pertenencia a cada clase de cada uno de los pares parámetro-valor que definen el caso que se desea clasificar.

- ***SMO*** implementa el algoritmo de optimización mínima secuencial (Platt, 1999, citado por Witten y Frank, 2005, p. 410) para entrenar una máquina de soporte vectorial (SVM) (Vapnik, 1995). Estos algoritmos extienden las características de los modelos lineales, ya que permiten distinguir entre clases que presentan límites de decisión no lineales. Para ello se transforman los datos originales transformándolos de forma no lineal en un nuevo espacio de mayor dimensión. En este nuevo espacio se construye un modelo lineal que pueda representar un límite de decisión no lineal en el espacio original. Puede hallarse una introducción más detallada a SVM en Burges (1998).

Algunos algoritmos se han completado con versiones *Bagging* o *Adaboosted*, que permiten mejorar los resultados aunque presentan mayor coste computacional (Duda et al., 2001).

- La técnica de ***Bagging*** —término derivado de *"bootstrap aggregation"*— utiliza múltiples versiones (de menor tamaño) del conjunto de entrenamiento y obtiene un clasificador para cada una. La clasificación final se alcanza por votación simple, ganando la clase que obtiene mayor coincidencia entre los clasificadores.
- La técnica de ***Adaboosting***(AB) —de *"adaptive boosting"*— se basa en la idea de obtener la clasificación final mediante una votación ponderada de diferentes clasificadores entrenados previamente con subconjuntos de muestras. En cada iteración las muestras se ponderan en función de si han sido correctamente clasificadas o no. La probabilidad de una muestra para ser utilizada por otro clasificador en una iteración posterior aumenta si está mal clasificada y disminuye en caso contrario. El peso de cada clasificador depende de su rendimiento en el conjunto de entrenamiento que se utilizó para construirlo.

4.6.1.3. Resultados

La tabla 4.5 resume los resultados del primer experimento de identificación automática según los diferentes algoritmos probados. En primer lugar se muestra el porcentaje global de identificación promedio de cada algoritmo (3ª columna). La estimación de la media se lleva a cabo promediando los resultados para cada conjunto de datos con un nivel de confianza de 0,95. Por lo tanto, los resultados se presentan en forma de intervalo de confianza para la media según la ecuación 4.2:

$$\mu = \bar{x} \pm 1{,}96 \cdot \frac{\sigma}{\sqrt{N}} \tag{4.2}$$

donde $\bar{x}$ y σ son, respectivamente, la media y la desviación estándar de los datos obtenidos en el experimento respectivamente y N el número de conjuntos de datos.

También se muestra el valor máximo obtenido con el conjunto de datos que figura entre paréntesis (4ª columna).

Tabla 4.5: Resultados más significativos de los algoritmos de aprendizaje automático utilizados para el experimento inicial de identificación de emociones.

Nombre	Descripción	Media(95 %IC)	Máx(Datos)
J48	Árbol de decisión con poda basado en C4.5	93,4 ± 2,1	96,4 (2G)
AB J48	Versión *Adaboosted* de J48	96,4 ± 1,5	98,3 (1L)
PART	Reglas de decisión basadas en árboles	94,2 ± 2,1	96,9 (2L)
AB PART	Versión *Adaboosted* de PART	96,7 ± 1,4	98,4 (1G)
DT	Tabla de decisión	88,7 ± 2,7	92,3 (1L)
AB DT	Versión *Adaboosted* de DT	93,4 ± 1,7	96,1 (1L)
IB1	Basado en instancias (1 solución)	93,3 ± 2,9	97,5 (2G)
IBk	Basado en instancias (k soluciones)	94,0 ± 2,4	97,9 (2G)
NB	*Naive Bayes* con discretización	94,6 ± 2,0	97,8 (1L)
SMO1	SVM con Kernel polinómico de 2^o grado	97,3 ± 1,3	99,0 (1G)
SMO2	SVM con Kernel polinómico de 3^{er} grado	97,1 ± 1,5	98,9 (1G)

Las dos versiones de SMO obtienen los mejores resultados tanto en el promedio como en el valor máximo de identificación. Las dos versiones *Adaboost* de J48 y PART obtienen resultados muy parecidos. Analizando los mejores resultados, se observa que SMO los obtiene con el conjunto Data1G, lo que muestra que la reducción basada en algoritmo genético (AG) supone una ayuda para estos sistemas, aunque las diferencias con Data1L y Data1LC no son significativas (véase la figura 4.8). Sin embargo, otros algoritmos (J48, IB1 e IBk) funcionan mejor con conjuntos de datos generados mediante un doble proceso de reducción de dimensionalidad (sin la 2ª derivada y aplicando, posteriormente, una reducción basada en GA). Finalmente se observa que existe un tercer grupo de algoritmos (Boost J48, PART, DT, Boost DT y NB) que mejoran si se elimina la redundancia que supone mantener las dos versiones, lineal y logarítmica, de F_0 y de la energía.

Cabe destacar también que Data1LC con menos de la mitad de los parámetros que Data1G o Data1L consigue prácticamente los mismos resultados, e incluso los mejora en el caso de los dos algoritmos basados en ejemplos —IB1 y IBk— (véase la parte superior de la figura 4.8). Un efecto similar ocurre con los conjuntos de datos sin la segunda derivada, ya que Data2LC consigue casi los mismos resultados que Data2G y Data2L con aproximadamente la mitad de atributos (parte inferior de la figura 4.8).

Por último, los resultados muestran que el hecho de eliminar información relativa al ritmo (Data1N y Data1NG) comporta una ligera disminución del porcentaje de identificación (entre 2 y 5 puntos). Sin embargo, los resultados empeoran significativamente si los parámetros se calculan únicamente en las vocales tónicas (Data1LS y Data2LS), disminuyendo entre un 5 % y un 12 % según el algoritmo de clasificación.

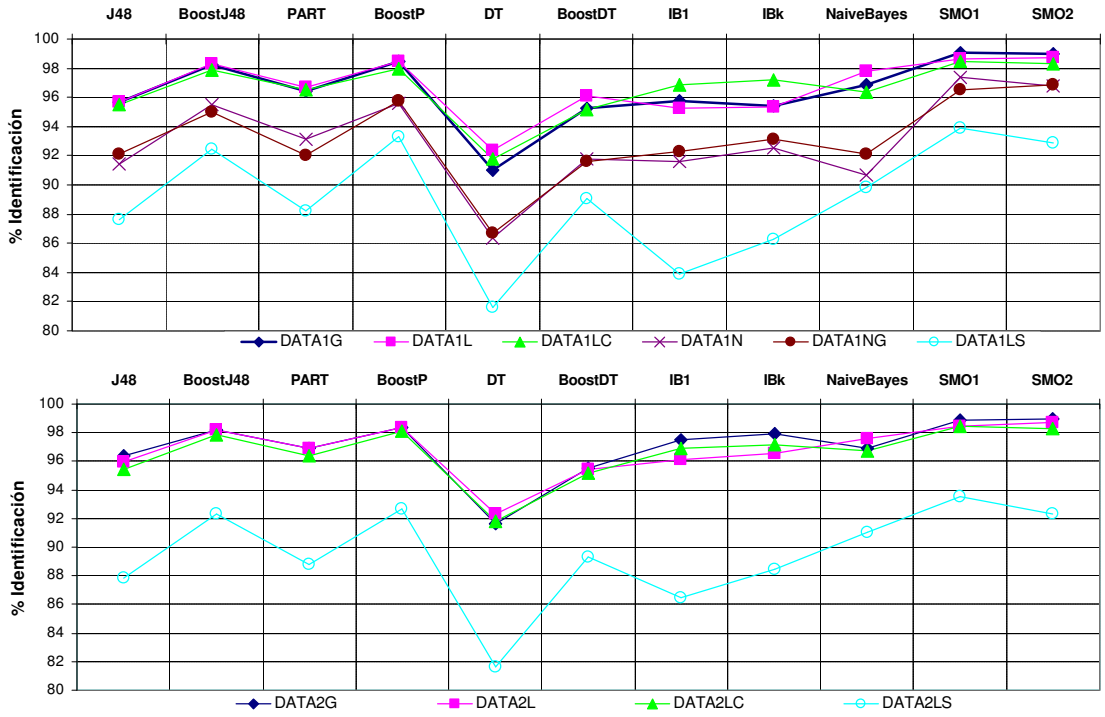


Figura 4.8: Porcentaje de identificación para cada algoritmo según el conjunto de datos.

La tabla 4.6 muestra la matriz de confusión con los resultados promediados para los once clasificadores con el conjunto de datos Data2G, que es el que ha conseguido el mejor porcentaje medio de identificación con un 97,02 % ± 1,23. La mayor confusión aparece entre los estilo neutro y sensual. También se produce una cierta confusión, aunque en menor porcentaje, entre los estilos alegre-agresivo y neutro-alegre. Si comparamos estos resultados con los del test subjetivo (véase la figura 4.5), podemos establecer un paralelismo desde un punto de vista cualitativo, ya que también se dan confusiones entre ambas parejas de estilos.

Si comparamos cuantitativamente los resultados, se observa que los participantes en la prueba subjetiva presentan un mayor porcentaje de confusiones en la pareja alegre-agresivo que en la pareja sensual-neutro, a diferencia de lo que sucede con el sistema automático. Una posible explicación es que subjetivamente se percibe alguna característica del habla sensual que no queda reflejada en los parámetros prosódicos utilizados en la identificación automática. Por ello, puede pensarse que se necesitará de la ayuda de parámetros relacionados con la cualidad de la voz para poder distinguir estilos prosódicamente parecidos pero acústicamente diferentes, como es el caso de la voz sensual, que presenta un tono parecido pero con una menor presencia de segmentos sonoros (voz susurrante).

La descripción de los procesos de producción y evaluación del corpus oral descrito se han aportado a la comunidad científica mediante su presentación en Iriondo et al. (2007b).

Tabla 4.6: Matriz de confusión promedio resultante del experimento de identificación automática con Data2G y los once clasificadores

Identificado →	AGR	ALE	TRI	NEU	SEN
AGR	**99,1 %**	0,8 %	0,1 %	0,0 %	0,0 %
ALE	1,6 %	**97,1 %**	0,0 %	1,2 %	0,2 %
TRI	0,2 %	0,1 %	**99,3 %**	0,4 %	0,1 %
NEU	0,2 %	0,9 %	0,4 %	**93,9 %**	4,5 %
SEN	0,0 %	0,1 %	0,2 %	4,9 %	**94,8 %**

4.6.2. Revisión automática guiada por los resultados del test subjetivo

Los resultados obtenidos en el apartado anterior serían excelentes desde el punto de vista del reconocimiento de la emoción en el habla, pero hay que aclarar que su objetivo consistía en validar la autenticidad del contenido expresivo simulado por la locutora profesional. Unas tasas tan altas de identificación automática únicamente indican que los estilos son lo bastante distintos para que los separe un clasificador automático. Analizando los resultados del test subjetivo, se ha constatado que no se alcanzan estos porcentajes de identificación, sino que existe un pequeño porcentaje de locuciones confusas o erróneas desde el punto de vista de la percepción de la expresividad. La confusión entre diferentes estilos, nos hace pensar que posiblemente los participantes en el test se han fijado en unas características del habla distintas a las que analiza el sistema automático.

Como consecuencia, surge la necesidad de sistematizar la eliminación de las locuciones con un contenido expresivo confuso o erróneo, de forma que sea una alternativa a una revisión manual de todas las locuciones del corpus por parte de un grupo de expertos; esta solución presentaría un primer problema debido al elevado coste de la revisión, unido a las dificultades en la coherencia de criterios entre evaluadores y a la nula reusabilidad de la metodología para el desarrollo de nuevos corpus. Por tanto, se plantea el diseño y el desarrollo de un sistema automático que mejore el contenido expresivo del corpus partiendo de las hipótesis siguientes:

- Los resultados del test subjetivo son más relevantes que los del experimento de clasificación automática, ya que el objetivo principal del corpus es la síntesis del habla expresiva y, por tanto, el material de partida para el modelado acústico posterior tiene que garantizar un mínimo de autenticidad desde el punto de vista de la percepción subjetiva.
- Es posible obtener un sistema de clasificación automático que simule la percepción subjetiva mayoritaria de un conjunto de evaluadores, mediante diferentes mejoras incorporadas al sistema (análisis acústico más completo, selección de atributos y combinación de clasificadores).

Por tanto, el objetivo del siguiente paso consiste en guiar el proceso de revisión automática según los resultados obtenidos en el test subjetivo, para conseguir finalmente una clasificación automática de las frases grabadas en dos categorías:

- Locuciones con un contenido expresivo significativamente parecido al estilo deseado. Estas locuciones tendrían que ser la mayoría, ya que se supone que la grabación la realiza un locutor o actor con los conocimientos y la experiencia suficientes.

- Locuciones confusas que no transmiten el estilo deseado, bien porque se confunden claramente con otro estilo de los del corpus, bien porque no se identifican con ninguno de los que contiene. Estas locuciones tendrían que ser la minoría y su posterior eliminación del corpus redundaría en una mayor calidad de éste.

4.6.2.1. Diseño del sistema propuesto

Se propone una solución basada en un clasificador óptimo (algoritmo/s y conjunto/s de atributos) capaz de modelar los criterios subjetivos obtenidos previamente en un test de percepción realizado con una parte relativamente pequeña, pero significativa, del corpus. Se pretende optimizar el proceso de clasificación automática para lograr la máxima coincidencia con los participantes del test subjetivo, es decir, que clasifique correctamente la clase (estilo) de las locuciones que tuvieron un alto porcentaje de identificación y no acierte la clase en las que presentaron una mayor confusión.

El esquema mostrado en la figura 4.9 resume la metodología seguida para revisar el contenido expresivo del corpus oral desarrollado. Destacan tres bloques fundamentales, que se describirán en los apartados siguientes, y que son:

- Realización de un test subjetivo con un conjunto de locuciones que sea suficientemente grande para ser representativo del resto del corpus pero suficientemente pequeño para que su realización tenga un coste de tiempo y de personal asumibles. Se requerirá una clasificación de las locuciones según el grado de expresividad percibido por los oyentes para su utilización en el proceso de revisión automática.

- La definición de una medida que permita comparar los resultados del test subjetivo con los del sistema automático, de forma que sirva de referencia para controlar los ajustes realizados en el sistema de clasificación.

- El desarrollo de un sistema de clasificación que permita ajustarse al máximo al criterio subjetivo, teniendo como referencia la medida de comparación establecida previamente.

4.6.2.2. Nivel de expresividad según el test subjetivo

Los resultados de un test subjetivo de clasificación se representan habitualmente mediante los porcentajes de identificación correcta y la matriz de confusión. Una de las entradas al sistema propuesto es el resultado del test subjetivo previo (véase el apartado 4.3). La forma habitual de representar dichos resultados es demasiado genérica para su utilización y, en consecuencia, se considera necesario representar el nivel de expresividad

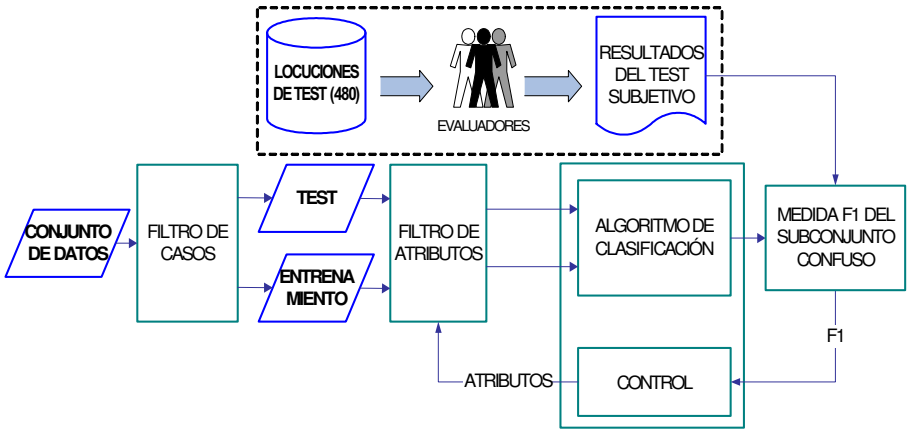


Figura 4.9: Diagrama de bloques de la revisión automática del contenido expresivo de las locuciones del corpus guiada por los resultados del test subjetivo

para cada frase del test de forma que este valor represente una medida de la calidad expresiva de la frase evaluada. Es decir, si una locución presenta un alto grado de confusión, su nivel de expresividad tiene que ser bajo; por el contrario, las locuciones con un alto porcentaje de identificación tendrán un nivel alto de expresividad.

Concretamente, la evaluación subjetiva de la expresividad del corpus se ha llevado a cabo sobre el 10 % de las locuciones del corpus aproximadamente (480 de 4.638). La matriz de confusión (véase la tabla de la figura 4.5, p. 56) muestra, por una parte, una ambigüedad clara entre los estilos alegre y agresivo y, por otra, cierta confusión entre sensual, neutro y triste. La respuesta *No lo sé / Otro* presentó, en general, una frecuencia de aparición baja, aunque aparece especialmente en los estilos neutro y sensual.

En la figura 4.10 se muestran dos histogramas basados en los porcentajes globales de identificación correcta (parte izquierda) y en la respuesta *No lo sé / Otro* (parte derecha). Por ejemplo, el histograma de la izquierda indica que 40 frases han sido correctamente identificadas por el 85 % de los oyentes que las evaluaron. Estos histogramas han permitido establecer dos simples reglas para decidir si una frase fue pronunciada de forma adecuada, desde el punto de vista de la expresividad, por la locutora profesional. Estas reglas eliminan los casos ambiguos que forman parte de la cola de los dos histogramas fijando dos umbrales: en el histograma de la izquierda el umbral se ha fijado al 50 % y, para el de la derecha, al 12 %.

Esto significa que las locuciones con un porcentaje de identificación correcta inferior al 50 % o con un porcentaje en la respuesta *No lo sé / Otro* superior al 12 % se consideran confusas desde el punto de vista de la expresividad percibida por los oyentes. La validación de estos dos umbrales se ha llevado a cabo volviendo a escuchar las frases que pasaban a considerarse confusas. Para el subconjunto de 480 locuciones utilizadas en la prueba subjetiva, existen 33 locuciones que no satisfacen como mínimo una de las dos reglas, lo que supone una eliminación del 6,88 % de los enunciados debida a la carencia de la expresividad adecuada.

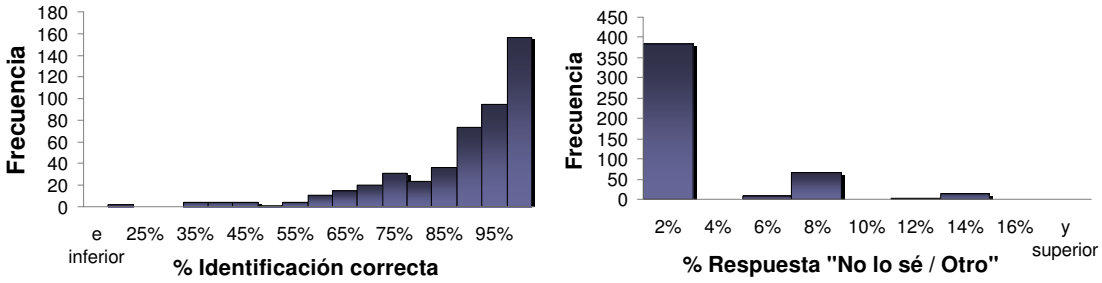


Figura 4.10: Histogramas del número de frases según el porcentaje de identificación correcta (izquierda) y el porcentaje en la respuesta *No lo sé / Otro* (derecha)

En resumen, las locuciones evaluadas en el test subjetivo se clasifican según las respuestas de los oyentes en dos clases, en función de si muestran una expresividad **significativa** o **confusa**.

4.6.2.3. Medida de comparación del nivel de expresividad

Los niveles de expresividad de las frases según el criterio subjetivo y según la clasificación automática tienen que representarse de la misma manera para posibilitar la definición de una medida de comparación entre ambos. Dado que se ha definido una clasificación discreta de la expresividad —en este caso, dos clases—, una medida adecuada para comparar ambos criterios es la F_1. La medida F_1 de una clase se calcula a partir de la precisión y la cobertura de la clasificación automática tomando como referencia, en nuestro caso, el resultado de la clasificación subjetiva (véanse las ecuaciones 4.3, 4.4, 4.5).

$$F1 = \frac{2 \cdot Precision \cdot Cobertura}{Precision + Cobertura} \tag{4.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$Cobertura = \frac{TP}{TP + FN} \tag{4.5}$$

donde TP (*true positive*) indica los casos clasificados correctamente en una clase, FP (*false positive*) los casos clasificados incorrectamente para esa clase y FN (*false negative*) los casos que se han dejado de clasificar correctamente. La medida F_1 pondera por igual precisión y cobertura, ya que es un caso particular de la medida F (ecuación 4.6) con el parámetro $\alpha = 1$:

$$F\alpha = \frac{(1 + \alpha) \cdot Precision \cdot Cobertura}{\alpha \cdot Precision + Cobertura} \tag{4.6}$$

donde α es un número real no negativo.

La medida F_1 siempre se sitúa entre los valores de cobertura y precisión, pero con tendencia hacia el menor de los dos. Las otras dos medidas comúnmente utilizadas son la

medida F_2, que pondera dos veces más la cobertura que la precisión, y la medida $F_{0.5}$, que pondera más la precisión.

En el presente trabajo únicamente se ha utilizado la medida F_1 para evaluar el resultado del clasificador automático según el etiquetado realizado a partir de las respuestas de los participantes en la prueba subjetiva que han evaluado 480 locuciones. Desde el punto de vista subjetivo, estas 480 locuciones se han divido en dos clases: **significativa** y **confusa** según el análisis de los resultados realizado en el apartado 4.6.2.2. Desde el punto de vista del clasificador automático, una locución se considerará **significativa** si la clasificación automática obtenida coincide con su estilo; o **confusa** si se produce un error en la clasificación.

4.6.2.4. Ajuste del sistema de clasificación automática

Por último, se requiere un método de ajuste del sistema de clasificación que, durante la fase de entrenamiento, pueda adaptarse según la evolución de la medida de comparación del nivel de expresividad. El objetivo final es conseguir que el sistema automático obtenga un criterio de evaluación de la expresividad lo más parecido posible al resultado del test subjetivo y que permita extrapolar este comportamiento a las locuciones no evaluadas en la prueba subjetiva.

Un posible ajuste del sistema de clasificación se puede llevar a cabo mediante la selección de los atributos que mejor representan las categorías establecidas. En los sistemas que utilizan un elevado número de atributos, una búsqueda exhaustiva de subconjuntos de atributos es muy costosa computacionalmente, y se puede optar por utilizar procedimientos de búsqueda del tipo *greedy* que garanticen encontrar un subconjunto de atributos localmente óptimo (Witten y Frank, 2005). Inicialmente, se han escogido, por un lado, el proceso de selección *Forward* (FW), que empieza sin ningún atributo y añade en cada iteración el más relevante; por otro lado, se ha desarrollado la técnica de eliminación *Backward* (BW), que parte del conjunto entero de atributos y elimina en cada iteración el menos significativo.

4.6.2.5. Resultados preliminares

Un primer experimento, descrito en Iriondo et al. (2007a), ha consistido en realizar seis pruebas combinando tres algoritmos de clasificación (SMO, Naive Bayes y J48; véase el apartado 4.6.1) con las dos técnicas de selección de atributos (FW y BW) comentadas en apartado anterior. Para cada algoritmo se ha realizado un proceso FW de forma que en cada iteración se ha aumentado en uno el número de atributos seleccionados, escogiendo el que consigue, junto a los atributos ya seleccionados anteriormente, un valor máximo de F_1 en la clase de frases confusas. Al tratarse de una clase binaria, maximizar la F_1 de una clase supone también maximizar F_1 de la otra clase. El proceso BW parte del conjunto entero de atributos y en cada iteración elimina el menos relevante, es decir, aquel que al ser excluido maximiza el valor de F_1. Una vez completado todo el ciclo FW (partiendo

del atributo localmente más relevante hasta haberlos incorporado todos) o todo el ciclo BW (partiendo del conjunto completo de atributos hasta quedarse sólo con uno), el mejor subconjunto de atributos será el que haya conseguido el valor máximo de F_1.

Para escoger el conjunto de atributos inicial, se ha partido del experimento de identificación automática del estilo/emoción presentado en el apartado 4.6.1, en el cual se ha analizado un amplio número de conjuntos de datos y algoritmos. Hay que recordar que, inicialmente, se partió de 464 atributos por frase leída, que se fueron reduciendo según diferentes estrategias (véase la figura 4.7). Los resultados obtenidos han mostrado que con el conjunto de datos Data2LC se logran muy buenos porcentajes de identificación con sólo 68 atributos. Por tanto, este conjunto de datos se ha escogido para el siguiente experimento de selección de los atributos que mejor permitan emular la percepción de los oyentes. Recordemos que este conjunto de datos representa la prosodia de cada locución mediante las secuencias (un valor por segmento) relativas a logF0, la energía en dB y las duraciones normalizadas mediante *z-score*. Se calcula también la primera derivada discreta de cada parámetro. Para las seis secuencias obtenidas se calculan las once medidas estadísticas enumeradas en el apartado 4.6.1.1. De esta forma, junto con los dos parámetros relacionados con el pausado, se toman en consideración un total de 68 atributos prosódicos por locución.

Para cada iteración del proceso de selección de atributos, se requiere un entrenamiento y un test del clasificador utilizado. Se trata de un entrenamiento supervisado en el que las clases que se deben predecir son los cinco estilos expresivos. Para esta tarea, el algoritmo utiliza las 480 locuciones de la prueba subjetiva como conjunto de test, y las 4.158 restantes como conjunto de entrenamiento. Además, las locuciones de test, 96 de cada estilo, están clasificadas en dos clases: **significativa** o **confusa**, según las dos reglas aplicadas a las respuestas de los oyentes en el test subjetivo previo (véase el apartado 4.6.2.2). En la fase de selección de atributos, para un subconjunto dado, se entrena el clasificador y se evalúa con las 480 frases de forma que, a partir del resultado de la clasificación automática, se les asigna una de las dos clases que miden la expresividad: **significativa** (la clasificación automática coincide con el estilo) o **confusa** (error en la clasificación). Por tanto, el funcionamiento del clasificador se evaluará en función de la medida F_1 del conjunto de la clase de frases confusas. Como se trata de una clasificación binaria, sólo hace falta fijarse en la medida F_1 de una de las clases, ya que los dos valores de F_1 tendrán el mismo comportamiento. Se ha escogido como referencia el valor F_1 de la clase con menos casos (33 de 480), ya que pequeños cambios en la clasificación llevan a variaciones apreciables de la F_1 de la clase minoritaria y menos significativas en la otra clase.

En la figura 4.11 se muestra la evolución del máximo de la medida F_1 según el subconjunto de atributos óptimo en cada iteración. El valor máximo de F_1 obtenido es de 0,5 para el algoritmo SMO con la estrategia BW de selección de atributos y un subconjunto de 15 ó 16 atributos, aunque la estrategia FW para este algoritmo obtiene casi el mismo valor, pero para un rango mayor del número de atributos. El resultado para NB también es similar con ambas estrategias, mientras que para J48 el resultado es mejor con FW que con BW. Además, SMO/FW es la configuración más estable, ya que consigue mantenerse en el máximo con un amplio número de subconjuntos de parámetros.

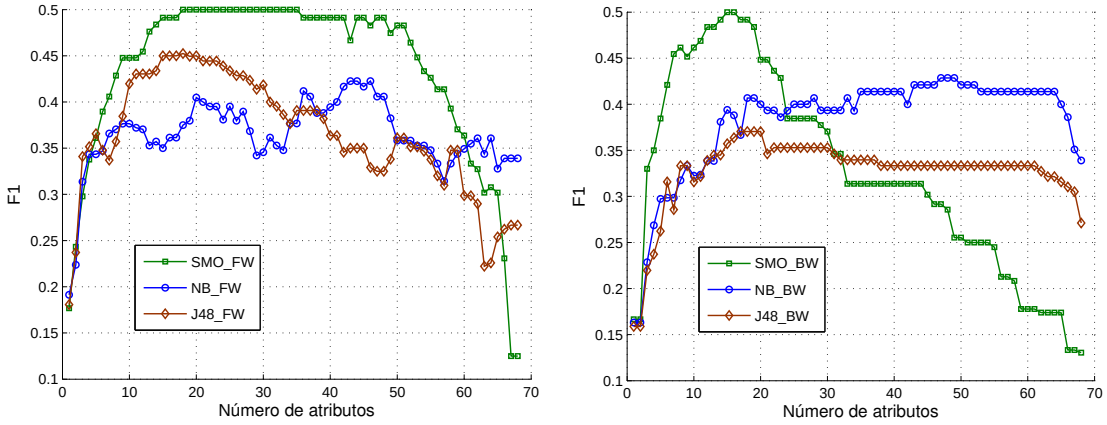


Figura 4.11: Valores máximos de F_1 para los algoritmos SMO, Naïve-Bayes y J48 con los subconjuntos de atributos obtenidos mediante: (a) selección *forward* y (b) eliminación *backward* partiendo del conjunto de datos Data2LC

En la tabla 4.7 se muestra para cada prueba el rango de número de atributos para los que se obtiene el valor máximo de F_1. Se muestran también los valores de precisión, es decir, el número de frases que serían eliminadas correctamente (siguiendo el criterio subjetivo) respecto el total de frases en las que la clasificación no coincide con su estilo preestablecido. La columna de la derecha muestra los valores de cobertura, es decir, el cociente entre el número de frases consideradas confusas por el sistema y las 33 frases consideradas confusas por los oyentes. Los valores mostrados en negrita corresponden al menor número de atributos. Esta tabla nos permite observar la existencia de configuraciones con mayor precisión y de otras con mayor cobertura. Aunque la configuración J48/FW es la que consigue un mayor número de coincidencias, su baja precisión (18/51=0,35) incide negativamente en el valor de F_1.

Tabla 4.7: Valores máximos de F_1 con la precisión y cobertura asociadas para cada combinación de algoritmo y estrategia de selección de atributos (FW o BW), indicando el rango de número de atributos para el máximo valor de F_1 (en negrita el mínimo número de atributos que obtiene dicho máximo).

Algoritmo/Estrategia	Núm. de atributos	F1 máx.	Precisión	Cobertura
SMO / FW	**18**-35	0,49	0,58 (14/24)	0,42 (14/33)
SMO / BW	**15**-16	0,50	0,56 (15/27)	0,45 (15/33)
NB / FW	**43**-44	0,42	0,39 (15/38)	0,45 (15/33)
NB / BW	**47**-49	0,43	0,52 (12/23)	0,36 (12/33)
J48 / FW	**18**	0,43	0,35 (18/51)	0,55 (18/33)
J48 / BW	**17**-20	0,36	0,45 (10/22)	0,30 (10/33)

Los valores de F_1 conseguidos hasta el momento nos indican que el sistema automático presenta un comportamiento similar a los usuarios en aproximadamente la mitad de las frases confusas. Se cree conveniente introducir algunas mejoras en los diferentes

elementos que componen el sistema con el fin de aumentar el valor de F_1. En el apartado siguiente se presenta una propuesta que incluye mejoras en diferentes procesos del sistema.

4.6.3. Mejoras y propuesta final del proceso de revisión automática

Los resultados obtenidos con el experimento preliminar no nos permiten dar la implementación por cerrada y aplicar la metodología de revisión a todo el corpus. En este apartado se presentan las soluciones aportadas para mejorar los resultados obtenidos hasta el momento. El análisis de las posibles causas de estos resultados insuficientes y las soluciones introducidas se indican a continuación y se detallan en los apartados siguientes:

1. **Parámetros de cualidad de la voz**. El conjunto de atributos utilizado hasta el momento se correspondía únicamente a parámetros prosódicos. Se ha observado que estos parámetros no son suficientes para discriminar algunos casos de locuciones que, en cambio, los usuarios no confunden. La escucha particular de estas locuciones nos lleva a la conclusión que hace falta incluir parámetros de cualidad vocal —del inglés *voice quality* (VoQ).

2. **Estrategia de selección de atributos**. Las estrategias FW y BW utilizadas de forma independiente no tienen la posibilidad de deshacer decisiones tomadas en iteraciones anteriores. Si se combinan ambas estrategias, es posible descartar alguna decisión previa y obtener un valor superior al máximo local que se hubiese conseguido.

3. **Combinación de clasificadores**. El experimento llevado a cabo muestra que existen algoritmos de clasificación más precisos y otros con mayor cobertura. También pueden presentar diferente comportamiento según el estilo de la frase. Por lo tanto, se puede mejorar el resultado final combinando las salidas de varios clasificadores (Witten y Frank, 2005).

4.6.3.1. Parámetros de cualidad de la voz

A pesar de incorporar el resultado del test subjetivo en el proceso de selección de atributos, la confusión del sistema automático entre los estilos sensual, neutro y triste ha sido superior a la mostrada por los oyentes. Los atributos prosódicos utilizados han resultado insuficientes para poder distinguir con precisión dichos estilos con los algoritmos de clasificación probados. Por lo tanto, la primera mejora que se plantea es la inclusión de parámetros de VoQ que permitan diferenciar mejor los estilos que presentan características prosódicas similares. Los parámetros de VoQ utilizados se calculan directamente de la señal de voz con el programa de análisis Praat[8], sin la necesidad de utilizar ningún tipo de transductor o *hardware* adicionales. Se ha partido de los parámetros propuestos por Drioli et al. (2003) y, basándonos en los resultados obtenidos por Monzo et al. (2007), el conjunto final que se ha utilizado es el siguiente:

[8] http://www.praat.org/

***Jitter*:** promedio de las diferencias en valor absoluto de periodos fundamentales consecutivos, dividido por el periodo medio del segmento analizado.

***Shimmer*:** promedio de la diferencia en valor absoluto de las amplitudes de periodos consecutivos, dividido por la amplitud media del segmento.

GNE (*Glottal-to-Noise Excitation Ratio*): cuantifica la relación entre la excitación debida a oscilaciones de las cuerdas vocales respecto la excitación producida por ruido turbulento (Michaelis et al., 1997). Comparándolo con otras medidas parecidas como el HNR (del inglés *harmonic-noise ratio*) o el NNE (del inglés *normalized noise energy*), es el único parámetro que se puede considerar prácticamente independiente del *Jitter* y del *Shimmer*.

HammI (*Hammarberg Index*): diferencia entre los máximos de energía de las bandas frecuenciales 0-2000 Hz y 2000-5000 Hz.

Do1000: aproximación lineal de la pendiente espectral por encima de 1000 Hz calculada por medio del método de los mínimos cuadrados.

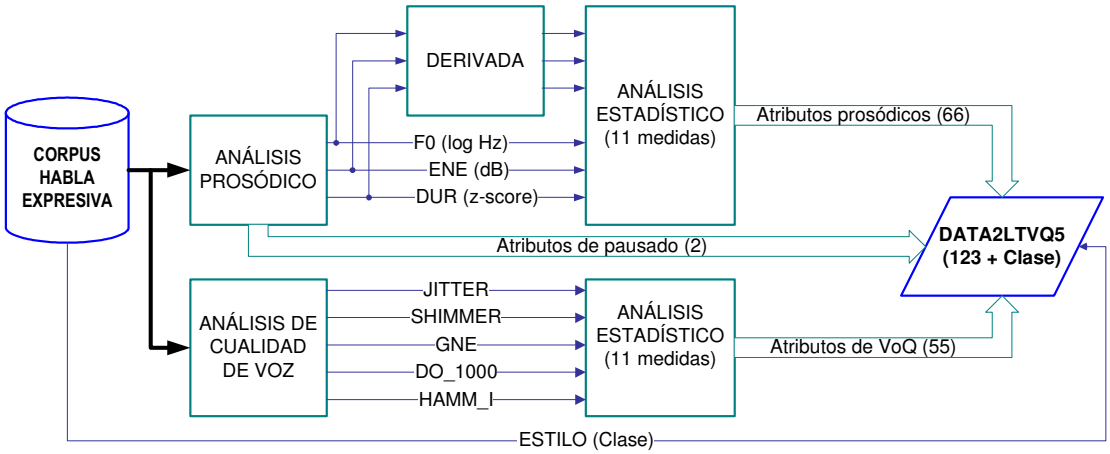


Figura 4.12: Generación del conjunto de datos para el sistema de validación final del corpus

Por tanto, el conjunto inicial de datos, que tenía 464 atributos prosódicos y que se había reducido a 68 debido a que los resultados de la clasificación automática fueron muy similares (véase el apartado 4.6.1), se complementa con las 11 medidas estadísticas aplicadas a las secuencias de los parámetros de VoQ, calculados únicamente en las vocales de la frase. La figura 4.12 esquematiza la generación del conjunto de datos definitivo, en el que se combina información prosódica e información sobre la VoQ, de modo que se obtienen 123 atributos para cada locución.

Los clasificadores entrenados con el nuevo conjunto de datos —que incluye atributos de VoQ— experimentan una mejora en el valor máximo de F_1. La figura 4.13 muestra la evolución del máximo de F_1 para los tres algoritmos SMO, J48 y NB con una estrategia de selección de atributos FW para este nuevo conjunto de datos, con valores máximos de $0,59$, $0,53$ y $0,48$ respectivamente. La tabla 4.8 muestra una comparación de los resultados

obtenidos mediante los 3 algoritmos estudiados (SMO, NB y J48), sin y con parámetros de VoQ. La incorporación de este tipo de parámetros supone una mejora absoluta de 0,1 para SMO, de 0, 09 para J48 y de 0, 06 para NB; en términos relativos, supone una mejora entre el 14 % y el 23 %. Con este nuevo conjunto de datos, no se ha probado la estrategia BW individual debido al elevado coste computacional que implica.

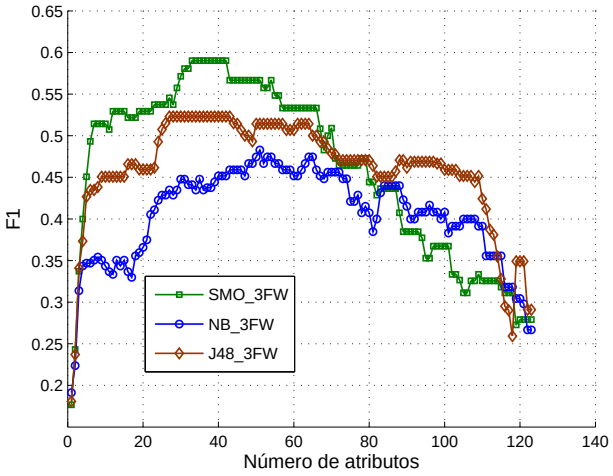


Figura 4.13: Valores máximos de F_1 por iteración para una estrategia de selección de atributos FW con el conjunto de datos que incorpora atributos de VoQ.

4.6.3.2. Selección de atributos FW-BW

Como se ha explicado en el apartado 4.6.2.4, se han desarrollado y probado dos técnicas *greedy*, una *forward* (FW) y otra *backward* (BW), para la selección de atributos. Si se combinan ambas estrategias de reducción de atributos, la solución final puede conseguir mejores resultados. Se ha programado un algoritmo que permite realizar p pasos hacia adelante (incorporando en cada paso el atributo más relevante) y q pasos hacia atrás (eliminando el menos relevante), siendo p y q dos números enteros positivos y que cumplen $p > q$. A esta estrategia la denominaremos, de aquí en adelante, pFW-qBW. Se ha escogido esta implementación y no otra realizada en sentido inverso por motivos de coste computacional.

La aplicación de una estrategia 3FW-1BW ha permitido mejorar los resultados de la F_1 máxima. Así, por ejemplo, la cuarta columna de la tabla 4.8 muestra los resultados obtenidos para los algoritmos SMO, J48 y NB, que mejoran en los 3 casos los resultados respecto a una estrategia FW simple. Los resultados más significativos se obtienen para J48 y NB, que consiguen una mejora absoluta de 0,06. En el caso de SMO la mejora absoluta es de 0,02. Hay que destacar que ambas modificaciones suponen una mejora relativa del máximo de F_1 en más del 20 % para los tres clasificadores. Respecto a la estrategia de 3FW-1BW, la estrategia 4FW-1BW iguala los resultados para SMO y J48, pero los empeora para NB.

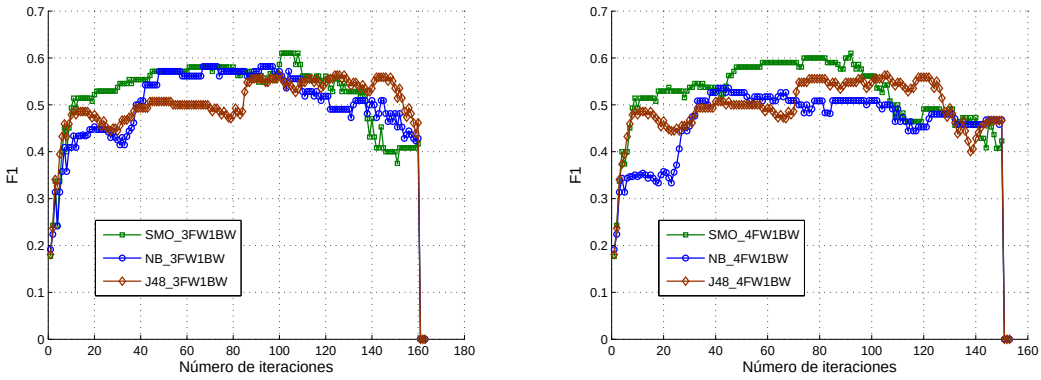
Tabla 4.8: Valor máximo de F_1 inicial con estrategia FW para los algoritmos SMO, J48 y NB, resultados con el conjunto de datos que incluye VoQ y, finalmente, con las estrategias 3FW-1BW y 4FW-1BW.

Algoritmo	Sin VoQ (FW)	Con VoQ (FW)	Con VoQ (3FW-1BW)	Con VoQ (4FW-1BW)
SMO	0,49	0,59	0,61	0,61
J48	0,43	0,52	0,56	0,56
NB	0,42	0,48	0,58	0,54

La figura 4.14 muestra la evolución del máximo de F_1 para los tres algoritmos SMO, J48 y NB con una estrategia combinada de selección de atributos 3FW-1BW. En cada iteración se escoge el subconjunto de parámetros que maximiza la medida F_1, añadiendo tres atributos en cada paso FW y eliminando uno en cada paso BW. El número de iteraciones (I) necesarias para completar el proceso viene dado por la fórmula:

$$I = N \cdot \frac{p+q}{p-q} - p \tag{4.7}$$

donde N es el número total de atributos; p y q son el número de pasos FW y BW respectivamente.



(a) (b)

Figura 4.14: Valores máximos de F_1 por iteración para el conjunto de datos que incorpora atributos de VoQ con las estrategias de selección de atributos: (a) 3FW-1BW y (b) 4FW-1BW.

4.6.3.3. Combinación de clasificadores

De los resultados del experimento descrito en el apartado 4.6.2, también se concluye que mientras que unos clasificadores son más precisos, otros ofrecen una mayor cobertura. Se pueden combinar múltiples modelos de diferentes tipos siguiendo un esquema de *stacking*[9] (Witten y Frank, 2005). Se trata de combinar las salidas de los diferentes cla-

[9]acrónimo de *stacked generalization*

sificadores con la finalidad de mejorar los resultados individuales. La versión más sencilla de *stacking* consiste en realizar una simple votación que puede ser ponderada o no. En cambio, hay técnicas de *stacking* más complejas que intentan aprender una serie de reglas que mejoren la clasificación individual. La figura 4.15 muestra el esquema de la técnica de *stacking*, en la cual se distinguen dos niveles de aprendizaje: el nivel 0, que corresponde a los clasificadores individuales, y el nivel 1, que consiste en un nuevo algoritmo de aprendizaje que tiene, como entrada, las salidas de los anteriores y, como salida, la nueva clasificación.

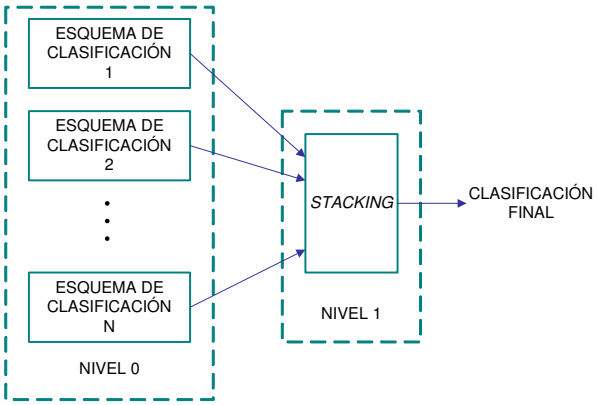


Figura 4.15: Combinación de diferentes clasificadores

Para la implementación de un sistema combinado, inicialmente se ha probado un simple sistema de votación con los siete mejores clasificadores seleccionados del conjunto de nueve clasificadores que se obtiene de combinar los tres algoritmos utilizados (SMO, NB y J48) con 3 estrategias de selección de atributos (FW, 3FW-1BW y 4FW-1BW). De los nueve clasificadores posibles se han descartado los dos que obtienen peor resultado: J48/FW y NB/FW (véase la tabla 4.8). Partiendo del pronóstico individual que clasificaría cada locución en expresivamente significativa o confusa, la decisión final se toma estableciendo un número mínimo de votos que la consideren mal interpretada. Analizando los resultados de los diferentes estilos, se ha observado que, en general, las frases confusas pertenecientes al estilo agresivo sólo eran detectadas por pocos clasificadores. Para mejorar el valor de F_1 resultante se ha realizado una votación ponderada para este estilo multiplicando por dos el número de votos recibidos. De esta forma, el estilo agresivo no quedaba tan penalizado al aumentar el número de votos mínimo para considerar confusa una frase. El valor máximo de F_1 es 0,71 que se obtiene con un mínimo de 4 votos (véase la figura 4.16) mejorando significativamente el mejor resultado individual, que es 0,61 (véase la tabla 4.8). Puede observarse que la secuencia de valores de la cobertura sigue una trayectoria decreciente a medida que aumenta el mínimo número de votos requeridos, mientras que la secuencia de la precisión es creciente.

Si en vez de un sistema de votación, se entrena otro algoritmo de *stacking* para el nivel 1 del esquema mostrado en la figura 4.15, se pueden mejorar un poco los resultados. Después de probar diferentes tipos de algoritmos con Weka, destacamos el resultado obtenido con el algoritmo PART (Witten y Frank, 2005), que se basa en obtener reglas

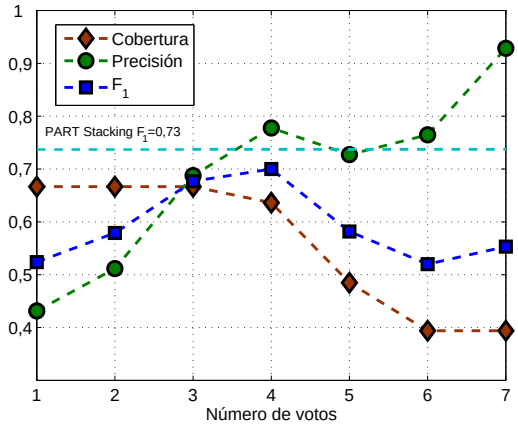


Figura 4.16: F_1, cobertura y precisión de la técnica por votación (adaptada con ponderación de 2 para los votos en el estilo agresivo) en función del mínimo consenso necesario para considerar las frases como confusas; se muestra también el resultado de F_1 obtenido con PART.

mediante árboles de decisión parciales.

El resultado con PART mejora ligeramente respecto la votación, consiguiéndose un valor de $F_1 = 0{,}73$. La reglas obtenidas con el entrenamiento se muestran en el algoritmo 2, donde el 0 significa clasificación coincidente con el estilo y el 1 clasificación de la frase en un estilo diferente al asignado a priori. Las dos clases de salida se refieren a expresividad SIGNIFICATIVA o CONFUSA (locución candidata a ser eliminada). Los clasificadores utilizados finalmente por PART después del entrenamiento son C1=SMO (3FW-1BW), C2=J48 (3FW-1BW), C3=J48 (4FW-1BW) y C4=NB (3FW-1BW). Para cada regla, los valores entre paréntesis de la derecha indican los casos bien clasificados y los mal clasificados, separados por una barra. El orden de las reglas es importante, ya que se aplica la primera regla que se cumple empezando desde arriba hacia abajo. La regla por defecto es la última de todas que asigna la clase mayoritaria. Hay que resaltar que el estilo agresivo tiene una regla específica (segunda línea) obtenida del entrenamiento. Esta particularidad ya se había considerado de forma heurística al incorporar un valor doble para este estilo en el sistema de votación ponderada.

Algoritmo 2 Algoritmo PART que implementa el nivel 1 de la estrategia de *stacking*.

```
C1 = 0 and C2 = 0 and C3 = 0 and C4 = 0: SIGNIFICATIVA (408/10)
C1 = 0 and Estilo = AGR and C2 = 1: CONFUSA (7/2)
C1 = 0: SIGNIFICATIVA (25)
C3 = 1: CONFUSA (18/3)
C2 = 0 and C4 = 0: SIGNIFICATIVA (4/1)
C2 = 0: CONFUSA (2)
: SIGNIFICATIVA
```

Finalmente, se muestra el número de frases eliminadas por estilo al aplicar la técnica combinada mediante votación (umbrales mínimos a 3 y 4 votos) o *stacking* con el algoritmo PART (véase la figura 4.17). La mejora de la medida F_1 (0,73) en el PART respecto la simple votación (0,71) se debe a un aumento en la precisión, ya que acaba seleccionando

menos frases para eliminar.

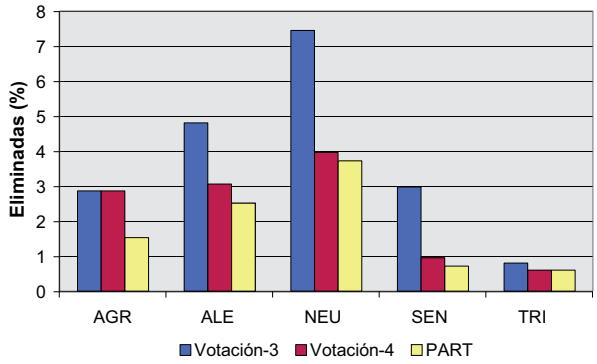


Figura 4.17: Locuciones eliminadas por estilo para las técnicas de *stacking* por votación (3 ó 4 mínimo número de votos) y PART (algoritmo 2).

La descripción del proceso final de revisión automática del corpus, que incorpora las mejoras introducidas en este apartado, se ha presentado en Iriondo et al. (2007c).

4.6.4. Evaluación del funcionamiento del sistema automático

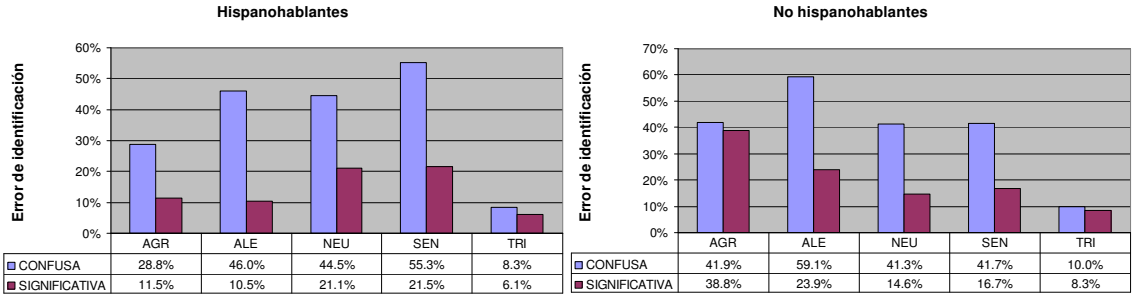
El proceso de revisión automática de todo el corpus ha dado lugar a un conjunto de locuciones consideradas confusas desde el punto de vista de su expresividad oral. Este resultado requiere una posterior validación que nos permita saber si el proceso es útil y, además, nos permita definir las acciones que deben realizarse con este subconjunto de frases confusas. De entrada, se pueden plantear diferentes acciones, que van desde eliminar automáticamente todas las locuciones propuestas por el sistema, hasta revisarlas de nuevo por una o más personas. Por lo tanto, se ha diseñado un nueva prueba subjetiva de identificación de emociones en la que se mezclarán las frases consideradas confusas por el sistema con otro conjunto de frases consideradas correctas.

La prueba está formada por las 75 frases que el sistema considera confusas en su versión de *stacking* con el algoritmo PART, distribuidas de la manera siguiente: agresivo (16 frases), alegre (23), neutro (24), sensual (6) y triste (6). Para formar el subconjunto de frases expresivamente significativas, se ha escogido el mismo número de frases que las consideradas confusas por el sistema para cada estilo. Como el conjunto de frases significativas es muy amplio, se ha optado por seleccionar frases que ya fueron evaluadas en el test presentado en el apartado 4.3 para garantizar que son representativas del estilo al que pertenecen. Se ha comprobado que estas frases también han sido clasificadas como significativas por el sistema. Por tanto, el test consta de 150 locuciones, que se presentan al oyente en orden aleatorio.

La hipótesis de partida de la prueba es que los sujetos cometerán más errores en las frases que el sistema automático ha clasificado como confusas (es decir, no las identificarán como portadoras de la emoción que pretendían reflejar), y sin embargo, identificarán correctamente las frases del otro subconjunto. La prueba la han realizado 38 evaluadores

hispanohablantes y 10 de lengua no hispana (dos italianos, dos franceses, tres irlandeses, dos austriacos y un finlandés) mediante la plataforma web ya utilizada en la prueba anterior. Se ha diseñado una evaluación de respuesta forzada a la pregunta: *"¿Qué estado emocional te transmite la voz de la locutora en esta frase?"*. De igual forma, las posibles respuestas son los 5 estilos del corpus más una opción *"No lo sé / Otro"*. Las personas de habla no hispana han utilizado una versión de la interfaz traducida al inglés. Los resultados de evaluadores que no entiendan el español nos permiten analizar si existe una alta dependencia de las respuestas con el contenido semántico de las frases.

Los resultados muestran claramente que los sujetos se equivocan mucho más en el conjunto de frases que el sistema considera confusas que en el otro. En la figura 4.18 se muestran los resultados correspondientes al error de identificación global por cada estilo para las dos clases: **confusa** y **significativa**. Una locución se considera erróneamente identificada por un oyente si la clasifica con un estilo diferente al que tiene asignado. Para los evaluadores hispanohablantes, se observa claramente un error de identificación muy superior para la clase **confusa** en todos los estilos a excepción del triste. Para el grupo de evaluadores de lengua no hispana, los resultados son muy parecidos a los del otro grupo para todos los estilos excepto para el agresivo, en el cual se produce casi la misma confusión para los dos tipos de frases. Para el estilo triste, el error de identificación es inferior al 10 % para todos los casos. Este hecho se debe a que este estilo se diferencia muy claramente de los demás, ya que posee unas características sonoras que hacen que prácticamente no se produzcan errores en su identificación. De hecho, en la primera prueba de validación que se realizó, este estilo obtuvo una identificación global del 98,8 % (véase el apartado 4.3).



(*a*) (*b*)

Figura 4.18: Porcentaje de error global de identificación subjetiva por cada estilo para las dos clases: **confusa** y **significativa**; según el grupo de oyentes sea: (a) hispanohablante o (b) de lengua no hispana

El primer análisis de los resultados, basado únicamente en porcentajes globales, sólo nos permite afirmar que el sistema se está comportando, en general, de la forma esperada, a excepción del estilo triste que, en este caso, no necesitaría una revisión posterior. Finalmente, es interesante presentar resultados considerando cada frase de forma individual. Para cada frase se dispone de la evaluación de cada oyente. En el apartado 4.6.2.2 se consideró que una locución pertenecía a la clase confusa si había obtenido un porcentaje de identificación inferior al 50 % o un porcentaje en la respuesta *No lo sé / Otro* superior al 12 % en la prueba de escucha realizada. Si aplicamos los mismos criterios en esta nueva prueba subjetiva, se obtienen los valores de F_1 mostrados en la tabla 4.9.

Puede observarse que el sistema se comporta de forma muy satisfactoria con los estilos sensual, alegre y neutro; y aceptable para el agresivo. El estilo triste no se ha evaluado de esta forma porque, con el presente criterio, ninguna frase se considera confusa. El resultado *Total* es la media ponderada de los cinco estilos según el número de locuciones de cada uno. En general, se observa una precisión muy alta y una cobertura menor. Este comportamiento indica que en la prueba realizada, el sistema automático ha detectado las locuciones confundidas por los oyentes, pero existe un cierto número de frases que la mayoría simple de usuarios identifican correctamente aunque el sistema las haya clasificado como confusas. Por lo que respecta al origen de los oyentes, la principal diferencia parece estar en la precisión, es decir, los evaluadores de lengua no hispana tienden a equivocarse más en las frases que pertenecen al grupo de las significativas que los hispanohablantes. Sin embargo, la cobertura presenta resultados parecidos para ambos grupos.

Tabla 4.9: Valores de precisión, cobertura y F_1 por estilo y global que indican la similitud de resultados del proceso de revisión automática y de la prueba subjetivo posterior para evaluadores hispanohablantes y de habla no hispana.

	Hispanohablantes			Habla no hispana		
Estilo	Precisión	Cobertura	F_1	Precisión	Cobertura	F_1
AGR	1,00	0,25	**0,40**	0,50	0,38	**0,43**
ALE	1,00	0,65	**0,79**	0,80	0,70	**0,74**
NEU	1,00	0,67	**0,80**	0,88	0,58	**0,70**
SEN	1,00	0,67	**0,80**	0,80	0,67	**0,73**
TRI	–	–	–	–	–	–
Total	1,00	0,57	**0,70**	0,76	0,58	**0,65**

4.7. Resumen

En este capítulo se ha explicado el proceso completo de producción de un corpus oral orientado a la síntesis expresiva del habla, desde su diseño hasta su validación final. Se ha contado con la colaboración del personal investigador del LAICOM-UAB en las tareas de definición de los estilos expresivos, su vinculación con textos publicitarios y la posterior grabación. La calidad del audio está garantizada por las condiciones en las que se ha realizado la grabación.

Sin embargo, al tratarse de habla expresiva obtenida mediante la lectura de textos semánticamente relacionados con los estilos definidos, se ha requerido de un proceso de validación que garantice que las locuciones que forman el corpus disponen del contenido expresivo adecuado. El corpus desarrollado tiene una duración de más de cinco horas dividido en cinco estilos expresivos: neutro, alegre, sensual, agresivo y triste. Una revisión exhaustiva de todo el corpus mediante pruebas de escucha sería excesivamente costosa.

En primer lugar, se realizó una prueba subjetiva de identificación de estilos sobre una muestra de aproximadamente el 10 % de las locuciones, junto con un experimento de idententificación automática de emociones en el habla (Iriondo et al., 2007b). La comparación de los resultados deparó comportamientos diferentes del sistema automático del criterio subjetivo general de los oyentes.

Entonces se propuso un método para ajustar, en la fase de entrenamiento, el sistema de identificación automática en función de los resultados obtenidos en la prueba de percepción dando lugar a un primer sistema que permitió la revisión completa del corpus de forma automática (Iriondo et al., 2007a).

Una serie de mejoras introducidas en diferentes módulos del sistema (selección de atributos *forward-backward*, inclusión de parámetros de cualidad de voz y combinación de clasificadores o *stacking*) consiguieron acercar más el funcionamiento del sistema al criterio subjetivo de los oyentes (Iriondo et al., 2007c) y, por lo tanto, generalizar la detección de aquellas locuciones que no se han pronunciado con la expresividad adecuada. Su eliminación o repetición permitirán un mejor modelado del habla expresiva y una base de datos de voz para la síntesis más adecuada.

Con los resultados de este último método aplicados sobre el corpus entero se ha realizado una segunda prueba subjetiva con 38 oyentes hispanohablantes y 10 de lengua no hispana, los resultados del cual han permitido validar su funcionamiento.

Capítulo 5

Modelado y estimación de la prosodia

El modelado prosódico tiene como principal objetivo determinar el comportamiento de los parámetros acústicos del habla asociados a una triple función: lingüística, extralingüística y paralingüística, generalmente en el nivel suprasegmental (véase el apartado 2.2.1). Los sistemas de CTH incorporan un módulo que permite estimar, a partir del texto, los valores de los parámetros acústicos que influyen principalmente en la percepción de la prosodia (apartado 2.3).

El objetivo perseguido en este punto de la tesis consiste en la obtención de un método para predecir los parámetros prosódicos en un sistema de síntesis del habla expresiva. Se pretende modelar de forma automática las funciones lingüística y paralingüística de diferentes estilos expresivos. El control de la función extralingüística de la prosodia no corresponde al ámbito de la presente tesis, ya que es un objetivo más propio de los trabajos de investigación relativos a la transformación del habla. Por tanto, la información extralingüística que aporte el modelado prosódico será inherente a las características del locutor utilizado.

El capítulo se inicia con una introducción que describe las primeras aportaciones del autor de la presente tesis al modelado prosódico para la síntesis del habla emocionada (apartado 5.1). A continuación, se presenta una propuesta de modelado y de predicción de la prosodia para la síntesis del habla expresiva basada en la aplicación de técnicas de aprendizaje automático al corpus de habla expresiva descrito en el capítulo 4 (apartado 5.2), la evaluación objetiva realizada (apartado 5.3) y, finalmente, una prueba de evaluación subjetiva del módulo desarrollado (apartado 5.4).

5.1. Primeras aproximaciones

En este apartado, se describen las dos primeras aproximaciones del autor de la presente tesis al modelado prosódico del habla expresiva orientado a la síntesis del habla. La experiencia obtenida durante la realización de estos dos trabajos ha servido de base para la definición de una parte importante de la presente tesis.

5.1.1. Modelado y validación de un modelo acústico de la expresión emocional en castellano

En el seno del LAICOM-UAB, Rodríguez et al. (1999) presentaron un modelo para la expresión emocional del habla en castellano que tenía como objetivo la mejora de la naturalidad en sistemas de CTH. Para este estudio, se partió de la hipótesis que el habla sufre cambios acústicos motivados directamente por las alteraciones fisiológicas que se producen en el cuerpo humano cuando un individuo experimenta una emoción y que dicha modificación depende de la lengua utilizada. A pesar de esta diferenciación, se consideró conveniente no distinguir entre procesos lingüísticos y no lingüísticos, considerando el habla emocionada como un sistema unitario que articula simultáneamente la influencia cultural de la lengua con los mecanismos fisiológicos de la emoción. El estudio se realizó a partir del análisis de formas sonoras suprasegmentales, ya que en ellas se combinan a la vez los caracteres propios de la lengua con los rasgos acústicos que determinan cada estado emocional.

La caracterización acústica del habla emocional se realizó mediante un análisis de la frecuencia fundamental, el contorno tonal, la duración segmental, la intensidad y el espectro. Según el enfoque de este estudio, la observación de segmentos muy cortos del habla no permite localizar y discriminar cuáles son los rasgos acústicos específicos de la emoción y qué influencia tiene una determinada lengua sobre ellos. Así pues, se determinó que la forma de garantizar un modelado eficaz de la expresión emocional del habla se tenía que basar en discursos orales completos.

El estudio realizado combinó una serie de pruebas previas de percepción con el análisis acústico del habla emocionada. El corpus de partida se constituyó a partir de la lectura de dos textos por parte de ocho actores —4 hombres y 4 mujeres— (véanse más detalles en el apartado 3.1.6.4). Mediante las pruebas de percepción se seleccionaron aquellos discursos que tuviesen todas las garantías de contener segmentos de habla con las emociones deseadas.

Se llevó a cabo un análisis sistemático de los 34 discursos seleccionados con el instrumento CSL-4300B de *Kay Elemetrics*. En aquellos casos en los que la detección de F_0 fue errónea se utilizó el analizador ANETO cedido por el *Grup de Tractament de la Parla* de la *Universitat Politècnica de Catalunya* (Febrer et al., 1998b). En dicho análisis se consideraron los parámetros indicados a continuación:

1. F_0: media, rango y variabilidad

2. Intensidad: media, rango y variabilidad.
3. Ritmo: duración total del discurso, tiempo de fonación, duración total de las pausas, número de grupos fónicos[1], número de pausas, duración de cada grupo fónico, duración de cada pausa, duración media del grupo fónico, duración media de la pausa, relación entre las pausas y la fonación, y número de sílabas por segundo.

Además de estos parámetros se añadieron dos tipos de representación gráfica:

1. Representación global de todo el discurso que incluía un oscilograma, una curva de F_0 y una curva de intensidad.
2. Representación parcial de segmentos que contenían entre uno y tres grupos fónicos mediante las mismas gráficas que la representación global.

Una vez realizado el análisis acústico, se necesitaba definir una medida de referencia con la que comparar los parámetros obtenidos. Por tanto, se definió el *estado-promedio* como la media aritmética de los datos de cada parámetro para cada locutor. De este modo, al poder cuantificar las desviaciones acústicas de cada voz respecto a su *estado-promedio*, se logró establecer un criterio común de referencia intralocutor que permitió comparar entre sí las voces de distintos locutores.

A partir del análisis realizado se observaron las siguientes características globales del habla emocionada:

- La estructura prosódica (contorno de F_0 e intensidad) característica de una emoción puede mostrarse únicamente en algunos grupos fónicos del discurso, siendo esta estructura parcial suficiente para que un oyente identifique el estado emocional.
- La estructura rítmica asociada a una emoción tiende a manifestarse a lo largo de la totalidad del discurso.
- La estructura melódica se caracteriza por una forma en diente de sierra. Los diferentes estados emocionales aumentan o disminuyen su variabilidad.

En referencia al modelo acústico asociado a cada emoción, se obtuvieron resultados válidos para seis de las siete emociones básicas estudiadas, ya que el asco se descartó por no superar el 50 % de identificación en la prueba subjetiva. En la tabla 5.1 se resumen los rasgos fundamentales de cada uno de los modelos acústico-emocionales.

Posteriormente, se llevó a cabo un experimento de validación del modelo descrito mediante síntesis del habla (Iriondo et al., 2000). La generación de habla emocionada mediante un sistema de CTH se realizó siguiendo los pasos que se enumeran a continuación:

- Construcción de un conjunto de frases portadoras con textos cuya información era semánticamente compatible con los distintos estados emocionales.

[1] Grupo fónico: porción del discurso comprendida entre dos pausas (Rodríguez et al., 1999).

Tabla 5.1: Resumen del modelo acústico de la expresión emocional para el castellano obtenido por Rodríguez et al. (1999) relativo al estado-promedio del locutor.

Alegría
Aumento del 10 al 50 % del tono medio
Aumento de la variabilidad tonal en un 120 %
Inflexiones tonales rápidas
Intensidad estable
Disminución del 20 % en la duración de las pausas
Deseo
Disminución del 10 % del tono medio
Disminución del 5 al 10 % de la variabilidad tonal
Inflexiones tonales lentas
Caída regular de la intensidad hasta 25 dB
Fuerte espiración al final de cada grupo fónico
Reducción de la duración de los grupos fónicos entre un 10 y un 20 %
Aumento de la fragmentación del discurso en un 20 %
Aumento del tiempo global del discurso
Rabia
Variación de la estructura tonal entre 20 y 80 Hz
Intensidad ascendente desde el inicio al final entre 5 y 10 dB
Reducción del número de pausas en un 25 %
Aumento de la duración de las pausas en un 8 % y del tiempo global del discurso
Aumento entre 10 y 15 dB en las bandas de 500-636 Hz y 2000-2500 Hz
Miedo
Aumento del 5 al 10 % del tono medio
Disminución del 5 % de la variabilidad tonal
Intensidad ascendente en 10 dB
Reducción de la duración de los grupos fónicos entre un 20 y un 25 %
Reducción de la duración de las pausas del 10 %
Sorpresa
Aumento del 10 al 15 % del tono medio
Aumento del 15 al 35 % de la variabilidad tonal
Grandes inflexiones tonales.
Aumento de la intensidad media entre 3 y 5 dB
Reducción de la duración de los grupos fónicos en un 10 %
Tristeza
Disminución del 10 al 30 % del tono medio
Disminución del 30 al 50 % de la variabilidad tonal
Ausencia de inflexiones tonales
Disminución de la intensidad media entre un 10 y un 25 %
Aumento de la fragmentación del discurso en un 10 %
Aumento de la duración de las pausas del 50 al 100 % (ralentización del discurso)

- Conversión a voz mediante EMOVS[2] de cada frase portadora, reproduciéndola tantas veces como modelos emocionales se quieren conseguir.
- Edición acústica de los parámetros prosódicos tomando como referencia los modelos obtenidos en el análisis del corpus de habla emocionada natural (véase el resumen presentado en la tabla 5.1). Las curvas de F_0 y energía y la duración de los segmentos de la frase se ajustaron siguiendo dicho modelo. Además, se hicieron algunos ajustes manuales guiados por los resultados de la síntesis con el objetivo de enfatizar la emoción deseada.

El trabajo de edición acústica mediante EMOVS reveló algunos aspectos nuevos e importantes, que no se habían detectado mediante el procedimiento previo de análisis acústico del habla natural:

1. La curva de entonación para algunas emociones presenta una forma en "diente de sierra" con pendientes de subida o bajada diferentes según la emoción expresada. La forma de este contorno se considera fundamental para el modelado acústico de las emociones.
2. El tipo de correspondencia en el tiempo entre los máximos de la energía y de F_0 también es un rasgo acústico determinante para caracterizar las expresiones emocionales.
3. Las relaciones entre la evolución temporal de la F_0 y de la energía, según sean cada una de ellas ascendente (A) o descendente (D), resultan relevantes para la expresión emocional, configurándose tanto relaciones F_0-energía directas (A-A y D-D), como inversas (A-D y D-A), según la emoción expresada.

A continuación se resumen las características de los contornos de energía y F_0 más importantes obtenidas en este primer trabajo para las cinco emociones modeladas: miedo, rabia, tristeza, alegría y deseo. Cabe recordar que el estudio se inició con un conjunto de siete emociones formado por estas cinco más la sorpresa y el asco. El asco se descartó debido a la baja tasa de identificación obtenida en la prueba de percepción realizada para validar el corpus grabado. En cambio, la sorpresa obtuvo una tasa de identificación suficiente para incluirla en el posterior análisis acústico y modelado, pero el intento de validación mediante la herramienta de síntesis del habla utilizada no fue satisfactorio.

Del **miedo** destaca que el contorno de F_0 en forma de "diente de sierra" tiene variaciones tonales muy rápidas (véase la figura 5.1). Los saltos bruscos y la estructura de meseta ascendente de la parte superior de cada "diente" hacen que el ataque tonal sea mucho más lento que la caída. Esta estructura asimétrica determina ese sonido característico de voz estrangulada que produce el miedo. Debe observarse como la energía es

[2]La herramienta gráfica EMOVS incluye un sintetizador de habla por concatenación de difonemas y trifonemas. La modificación de la curva de F_0 y de la duración de los segmentos se realiza mediante el proceso de interpolación de tramas descrito en Iriondo et al. (2003), trabajo en el que se utilizaba un análisis y una síntesis del habla similar a la técnica TD-PSOLA (Moulines y Charpentier, 1990). La variación de intensidad se consigue aumentando o disminuyendo la amplitud de la señal de voz.

globalmente ascendente y sus ascensos y descensos son coherentes y sincrónicos con los de F_0 (relación A-A y D-D).

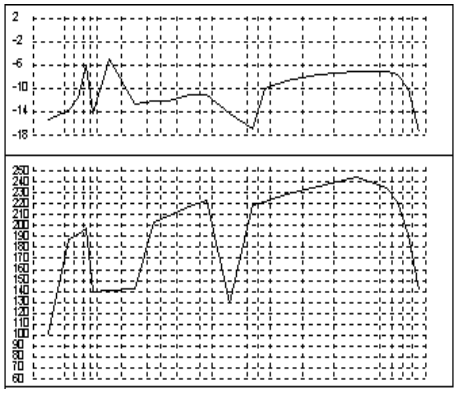


Figura 5.1: Ejemplo de los contornos de energía y de F_0 para una frase generada con los patrones definidos para el miedo.

En la **rabia**, la variabilidad del "diente de sierra" de la curva de F_0 es prácticamente la misma que para el miedo y la relación F_0-energía es también sincrónica y con una relación A-A y D-D. La diferencia entre el sonido del miedo y el de la rabia depende esencialmente de la simetría inversa que tienen los "dientes" (figura 5.2). En la rabia el ataque tonal es mucho más rápido que la caída; la meseta del "diente" es descendente. También es relevante que mientras en el miedo los "dientes" son más anchos (bi o tri silábicos), en el caso de la rabia tienden a ser estrechos (monosilábicos). La violenta y repetida subida tonal asociada a máximos de energía genera esa sensación característica de sucesión de golpes furiosos que tiene una locución con rabia.

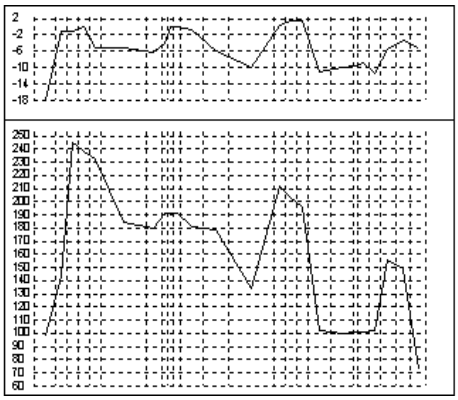


Figura 5.2: Ejemplo de los contornos de energía y de F_0 para una frase generada con los patrones definidos para la rabia.

En cambio, para la **tristeza**, la variabilidad, tanto en la F_0 como en la energía, es mínima. Mientras la variabilidad tonal en el miedo o la rabia puede rebasar los 140 Hz, en el caso de la tristeza no va más allá de 30 Hz, iniciándose el descenso ya desde un tono muy bajo. La estructura en "diente de sierra" no llega a configurarse, lo que provoca una baja variabilidad tonal en los segmentos, contrariamente a lo que sucede en el miedo o en

la rabia, produciéndose ese efecto de monotonía y lentitud tan característico de la tristeza. La relación entre F_0 y energía sigue siendo coherente, y ambas presentan, globalmente, un perfil descendente (figura 5.3).

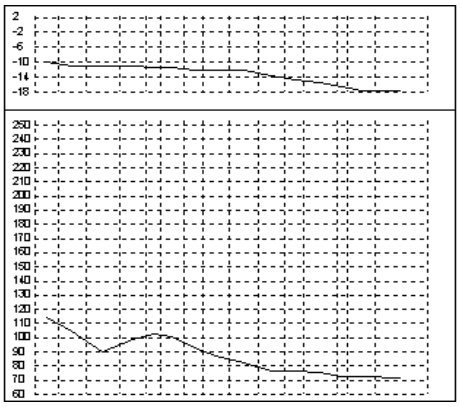


Figura 5.3: Ejemplo de los contornos de energía y de F_0 para una frase generada con los patrones definidos para la tristeza.

En el caso de la **alegría** la variabilidad tonal es también muy grande, como en el miedo y la rabia; no obstante, el "diente de sierra" de la alegría es simétrico, siendo los tiempos de ataque y de caída tonal muy similares. Otra diferencia importante es que el tono no se mantiene estable en el punto de máxima tensión formando una meseta, sino que desciende enseguida. Sin embargo, el rasgo acústico más relevante es el tipo de correspondencia entre los máximos de F_0 y los de la energía. Como puede observarse en la figura 5.4, energía y F_0 no son sincrónicos.

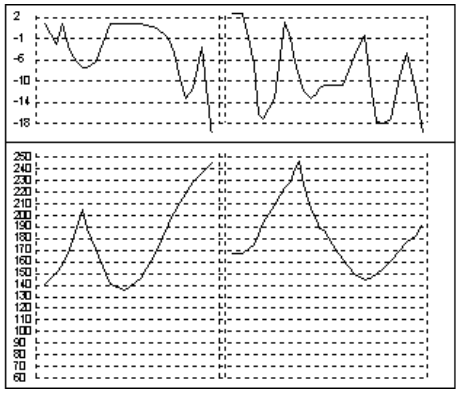


Figura 5.4: Ejemplo de los contornos de energía y de F_0 para una frase generada con los patrones definidos para la alegría.

En el **deseo**, la estructura del diente de sierra tonal vuelve a ser simétrica (tiempo de ataque similar al de la caída tonal), aunque su variabilidad es menor y sus puntos de inflexión son mucho más suaves, sin cambios bruscos. En esta emoción nos encontramos de nuevo con un desfase temporal entre la evolución de F_0 y la de la energía, en este caso mucho mayor que en la alegría. De hecho, en el deseo podemos hablar de una estructura

invertida de tono e intensidad, en la que máximos y mínimos tienden a ser opuestos y la evolución tiende a ser A-D o D-A de manera sistemática (figura 5.5). La suavidad de las evoluciones en el tiempo y su estructura opuesta a la del estilo agresivo (cuya sincronía entre intensidad y tono es muy precisa) es lo que da al habla ese sonido dulce y sensual que caracteriza a una voz que intenta seducir.

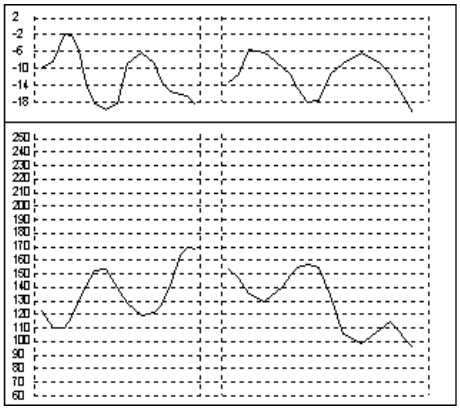


Figura 5.5: Ejemplo de los contornos de energía y de F_0 para una frase generada con los patrones definidos para el deseo.

Las muestras de habla generada únicamente se evaluaron de manera informal. No se realizó ninguna prueba de percepción con el número de frases y oyentes necesario para poder obtener unos resultados concluyentes. Por lo tanto, a partir de las muestras de habla emocional sintetizadas, se constató que la modificación de los parámetros prosódicos siguiendo los modelos acústicos del miedo, la rabia y la tristeza podían ser adecuados para integrarse en un sistema de síntesis concatenativa como el utilizado. En cambio, para conseguir los modelos definitivos del deseo y la alegría se consideró que era necesario mejorar el modelado acústico y, además, incorporar nuevas prestaciones al sintetizador: mayor capacidad de modificar la prosodia y control de los parámetros de cualidad de la voz.

5.1.2. Adaptación del modelo prosódico al catalán

La realización del paso siguiente trató de mejorar la capacidad expresiva del sistema de CTH en catalán mediante la inclusión de unas reglas simples para la modificación de los parámetros prosódicos. Cabe destacar que se trató de un experimento de ingeniería con el que se buscó obtener resultados de forma rápida. Este experimento forma parte, junto con el que se ha detallado en el apartado anterior, de la primera aproximación a la síntesis del habla emocional utilizando los recursos disponibles en aquel momento y que, como se constató posteriormente, resultaron insuficientes. Además de la falta de recursos, estos primeros experimentos también carecieron del rigor lingüístico necesario para una investigación en este ámbito. En concreto, para este experimento se utilizaron muestras de habla emocionada en castellano, de las que se extrajeron los valores de los parámetros prosódicos que permitía modificar el sistema de CTH en catalán para generar una colección de muestras de habla emocionada en esta lengua. Se han escogido las cuatro emociones

que tienen una expresión más universal (véase el apartado 2.1.3.1) con la intención de minimizar las diferencias asociadas al cambio de lengua.

Por lo tanto, la descripción que se realiza a continuación tiene como objetivo ilustrar el punto de partida del presente trabajo.

Se llevó a cabo la generación e implementación de un modelo prosódico específico para cada una de las cuatro emociones consideradas las más básicas: miedo, rabia, tristeza y alegría. De estas cuatro emociones, las tres primeras habían obtenido resultados prometedores mediante el modelado prosódico y, en cambio, la alegría presentó mayores dificultades. Como se describe más adelante, se realizó una validación subjetiva del método propuesto que corroboró este comportamiento también para esta aproximación.

La metodología seguida, teniendo en cuenta que no se disponía de un corpus de habla emocionada para el catalán, fue la siguiente (véase el esquema presentado en la figura 5.6):

1. Se parte del corpus de habla emocionada para el español descrito en el apartado 5.1.1. De dicho corpus se escogen 4 locuciones correspondientes al mismo texto con las 4 emociones que se desea modelar. Cada locución está formada por 7 frases que se segmentan en fonemas o alófonos y se etiquetan con sus valores de F_0 media, energía media y duración. Se anotan también la duración de las pausas entre frases.

2. Se traduce el texto al catalán y, una vez obtenida la transcripción fonética de forma automática con la herramienta descrita en el apartado E.1, se asignan los valores de los parámetros prosódicos de cada segmento a partir de la información almacenada en la base de datos para el castellano mediante un alineamiento temporal de las secuencias de valores prosódicos en el nivel segmental.

3. Se sintetiza habla emocionada a partir del texto traducido al catalán y de la información prosódica ajustada manualmente.

4. Se realiza una prueba de percepción para validar que el habla sintetizada incorpora la expresión deseada.

5. Se obtiene un modelo de modificación prosódica respecto a la salida por defecto del sistema de CTH en catalán (expresividad neutra).

6. Por último, se automatiza dicho modelo, incorporándolo al módulo de PLN del sistema de CTH.

7. Se realiza una segunda prueba subjetiva de identificación de emociones con el fin de valorar los resultados que ofrece el sistema automático.

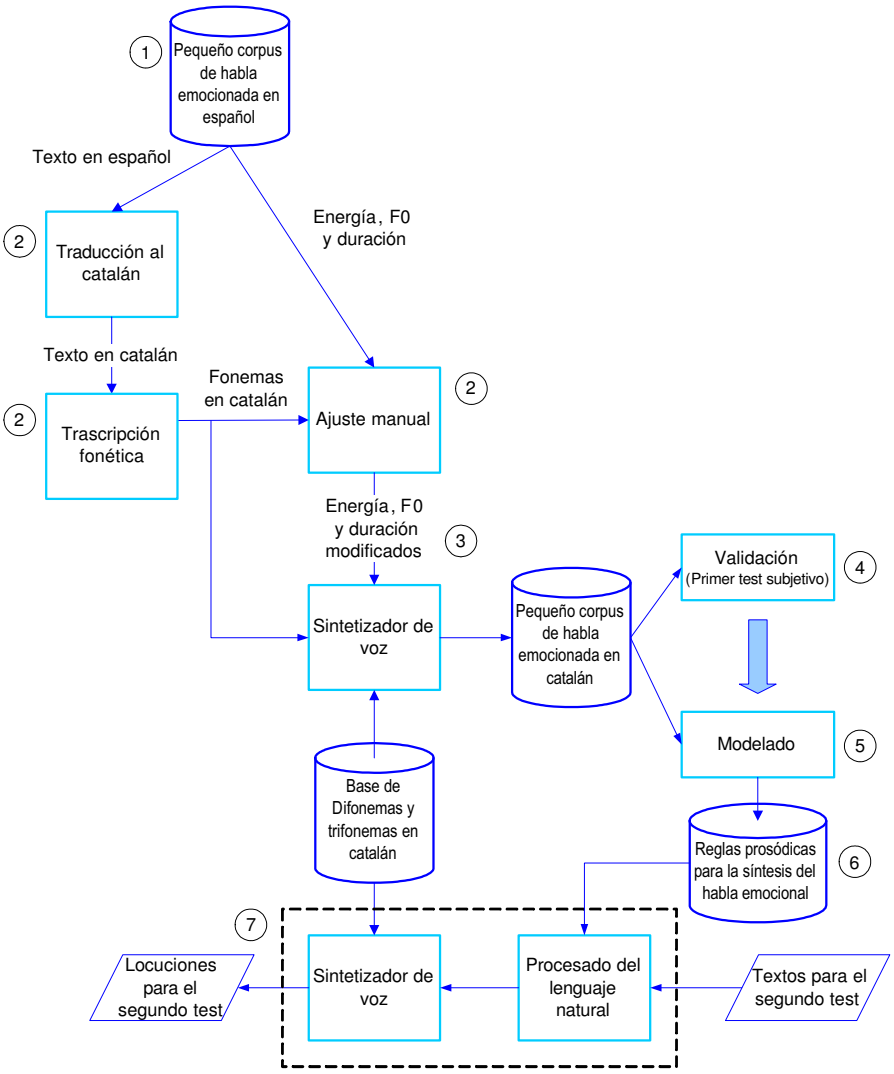


Figura 5.6: Diagrama de bloques que resume los siete pasos seguidos durante los procesos de definición y validación del modelo prosódico orientado a la síntesis del habla emocional en catalán.

El texto original en castellano que sirvió para la grabación de la muestra de habla emocionada utilizada en este experimento es el siguiente:

"La casa apareció al final del camino. Empezaba a ponerse el sol, pero la fachada del edificio aun se veía con claridad. Unas figuras pasaban por detrás de las ventanas del piso superior. Me acerqué poco a poco. Nadie me vio, nadie me esperaba, nadie me recibió, entré sin hacer ruido. Subí las escaleras con agilidad. Las voces me guiaron hasta la gran habitación y lo vi todo."

La traducción al catalán que sirvió para generar una muestra de habla sintética emocionada es la siguiente:

"La casa aparegué al final del camí. Començava la posta de sol, però la façana de l'edifici encara es veia amb claredat. Unes figures passaven per darrera de les finestres del pis superior. Em vaig apropar a poc a poc, ningú em veié, ningú m'esperava, ningú em rebé. Vaig entrar sense fer soroll. Vaig pujar les escales amb agilitat. Les veus em guiaren fins a la gran habitació i ho vaig veure tot."

A continuación se detallan las tres fases principales de esta aproximación: *i)* la generación de un corpus de habla emocionada en catalán mediante técnicas de síntesis del habla; *ii)* la generación de un modelo prosódico sencillo para habla emocionada en catalán y *iii)* la incorporación de dicho modelo en el sintetizador y la evaluación subjetiva de los resultados.

5.1.2.1. Generación de la muestra de habla emocionada en catalán

La colección de frases de habla emocionada en catalán se generó con el sistema de CTH a partir de una entrada formada por la transcripción fonética y la información prosódica ajustada manualmente para las cuatro emociones estudiadas. La transcripción fonética de los textos originales traducidos al catalán se obtuvo de forma automática con la herramienta descrita en el apartado E.1. La información prosódica asociada a cada segmento se calculó mediante el encaje de los valores prosódicos de los segmentos de las locuciones del castellano en la cadena de segmentos de las respectivas frases en catalán. La energía y la F_0 se asignaron mediante un alineamiento temporal de sus contornos. Las duraciones de las pausas se copiaron directamente. La duración de las frases se ajustó globalmente; se alargaron o acortaron los segmentos proporcionalmente para cada emoción. Las duraciones individuales de cada segmento se multiplicaron por la relación entre la duración de la frase original con la emoción deseada y el promedio de las duraciones de la misma frase con las cuatro emociones.

Como resultado, se obtuvo un pequeño corpus de habla sintética masculina en catalán para cuatro emociones correspondientes al mismo texto. Además, se generó una locución del mismo texto con el sistema de CTH en su modo por defecto (que denominaremos expresión "neutra") y que sirvió como patrón de comparación.

La validación del corpus de habla sintética emocionada se llevó a cabo mediante una prueba perceptiva realizada con diez oyentes que evaluaron en dos fases los estímulos generados. Antes de empezar la prueba, se dio la oportunidad al oyente de escuchar una voz sintética neutra, para que este se familiarizase con el sonido del sintetizador del habla.

En la primera fase, el objetivo era disponer de una estadística de la primera impresión que producían los estímulos correspondientes a cada emoción sin poder establecer un criterio comparativo entre todas ellas. En la segunda se quiso valorar si el reconocimiento aumentaba una vez ya se habían escuchado las diferentes emociones y, por tanto, ya existía un criterio comparativo entre ellas. El orden de las locuciones presentadas fue aleatorio en ambas pruebas con el fin de que el oyente evaluador no se ayudara del resultado de la primera fase.

En las dos pruebas se pidió a los oyentes que eligiesen, después de escuchar cada locución, entre las siguientes posibilidades: miedo, alegría, rabia, tristeza o emoción no identificada. El tanto por ciento de reconocimiento acústico de emociones en las dos fases de la prueba se puede observar en la figura 5.7. La tristeza es la emoción con mayor índice de reconocimiento en las dos fases, llegando a un 100 % de reconocimiento en la segunda. La rabia y el miedo mejoran sus porcentajes una vez ya se han escuchado todas las emociones, gracias a que es más fácil establecer diferencias entre ellas. La alegría es la única emoción que empeora los resultados del reconocimiento, pasando de un 40 % a un 30 %. Estos resultados coinciden con la premisa de la que ya se partía: la alegría es la emoción más difícil de modelar mediante parámetros puramente prosódicos. Además, algunos oyentes pusieron de manifiesto la dificultad de reconocerla debido al contenido semántico del texto, que les pareció más terrorífico que alegre.

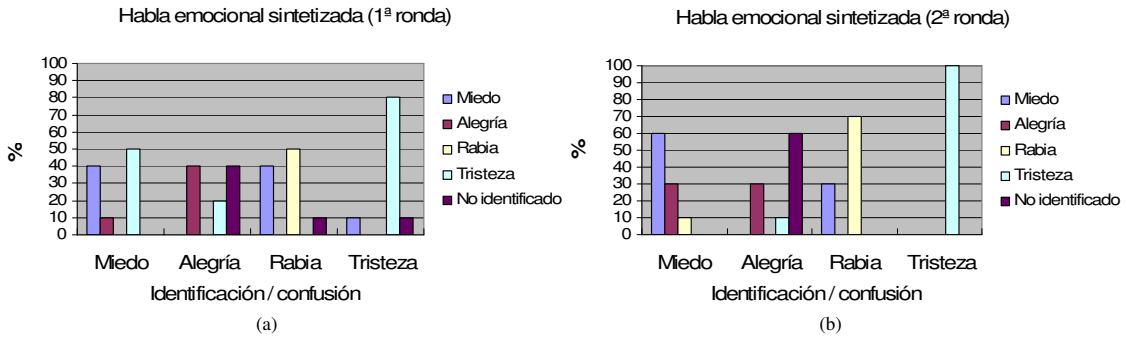


Figura 5.7: Porcentajes de identificación de las cuatro emociones en el test perceptivo realizado con locuciones sintetizadas obtenidas a partir de un ajuste manual de la prosodia.

5.1.2.2. Modelo prosódico para habla expresiva en catalán

El principal objetivo de este modelado prosódico fue obtener datos sobre el comportamiento general de los principales parámetros prosódicos relacionados para las cuatro emociones consideradas, con la intención de incorporarlos al sistema de síntesis concatenativa existente en ese momento. Se consideraron tres categorías de parámetros prosódicos: F_0, ritmo y energía.

F_0. Se calcula el valor medio y la variación (diferencia entre el valor máximo y el mínimo) de F_0 para cada frase. En la tabla 5.2 se muestra la variación promedio de estos dos parámetros en tanto por ciento para cada emoción. Según estos resultados, el miedo se caracteriza por una subida muy alta del tono medio y, prácticamente, por la misma variabilidad que el estilo neutro. La rabia presenta un aumento importante tanto en el tono medio como en su variabilidad. La tristeza muestra un pequeño descenso del tono medio y una disminución muy acusada de la variabilidad. Sin embargo, en la alegría no se encuentra una modificación significativa de estos dos parámetros. En la figura 5.8 se muestra la media y la desviación estándar del tono medio y de su

variación a partir del análisis de las frases que componen la base de datos generada de forma sintética.

Tabla 5.2: Porcentaje relativo de variación de los parámetros de F_0 con respecto al estilo neutro para cada emoción.

Variación relativa	Miedo	Alegría	Rabia	Tristeza
F_0 media	+52 %	+13 %	+33 %	-7 %
Variación de F_0	-3 %	-10 %	+30 %	-60 %

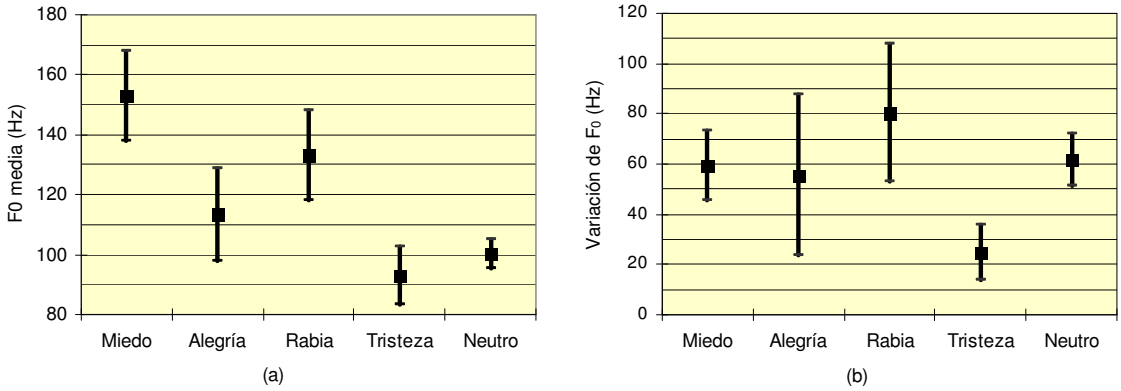


Figura 5.8: Media y desviación estándar del promedio de F_0 (a). Media y desviación estándar de la variación de F_0(b).

Ritmo. El ritmo del discurso se caracteriza en este trabajo mediante la duración de las pausas y la duración de los grupos fónicos. Para el estudio sobre la duración de las pausas se optó por calcular el valor medio de la duración de las pausas de las cuatro emociones y de la neutra. Una vez calculados, se determinaron los incrementos o decrementos medios respecto a la versión neutra (véase la tabla 5.3).

Respecto a la duración de los grupos fónicos, se calcula el porcentaje relativo de las duraciones de los segmentos de una frase con emoción respecto los del estilo neutro. Generalmente se observa que el final de la frase suele presentar un patrón de duración segmental diferente al del resto de la frase, consecuencia de su posición prepausal. Por este hecho, el último grupo fónico se estudia de manera distinta a la del resto de la frase. La tabla 5.4 muestra que la velocidad del habla aumenta para la rabia y disminuye claramente para la tristeza.

Tabla 5.3: Porcentaje relativo de variación de la duración media de las pausas respecto al estilo neutro

Variación relativa	Miedo	Alegría	Rabia	Tristeza
Duración	+38 %	+3 %	-13 %	+128 %

Tabla 5.4: Porcentaje relativo de variación de la duración media de los grupos fónicos respecto al estilo neutro

Variación relativa	Miedo	Alegría	Rabia	Tristeza
Duración del último grupo fónico	+9 %	+0.2 %	-7 %	+25 %
Duración del resto de grupos fónicos	+6 %	+12 %	-4 %	+23 %

Energía. Los parámetros de energía describen características de la amplitud de la señal de voz. Se ha calculado la energía media y la variación de cada frase del corpus. En la tabla 5.5 se presentan los resultados calculados como incrementos o decrementos de la energía media y la variación respecto el estilo neutro. Esta variación se expresa en decibelios (dB). Se observa que el estilo neutro presenta una menor variación de energía, ya que para las cuatro emociones se produce un incremento del valor de este parámetro.

Tabla 5.5: Variación relativa de los parámetros de energía respecto al estilo neutro en dB

Variación relativa	Miedo	Alegría	Rabia	Tristeza
Energía media	-0,16 %	+0,29 %	+1,13 %	-1,46 %
Variación de energía	+13 %	+11,1 %	+14,3 %	+10,4 %

5.1.2.3. Automatización y evaluación del modelo prosódico

La automatización del modelo prosódico consistió en la definición de un conjunto de reglas que representan los resultados obtenidos con el análisis previo. Estas reglas se definen como una modificación de los parámetros prosódicos generados automáticamente por el sistema de CTH para el estilo neutro. El tono y la energía están representados por dos parámetros (media y variación) cada uno. El ajuste de los valores se realiza para cada segmento de la frase según los pasos siguientes:

1. A partir del texto, los valores de los parámetros prosódicos se calculan para cada segmento (estilo neutro), p_0 en las ecuaciones (5.1) y (5.2).
2. Se normaliza p_0, restando el valor medio $\bar{p}_0$
3. Se ajustan los valores normalizados a la variación deseada siguiendo la ecuación (5.1), donde $\Delta\bar{R}$ es el factor de corrección de la variación.
4. Los valores finales, p_f, se obtienen añadiendo el nuevo valor medio a los valores calculados en el paso 3. En la ecuación (5.2), $\Delta\bar{A}$ es el factor de corrección de la media.

$$\hat{p} = \Delta\bar{R} \cdot (p_0 - \bar{p}_0) \tag{5.1}$$

$$p_f = \hat{p} + \Delta\bar{A} \cdot \bar{p}_0 \tag{5.2}$$

Los factores de corrección $\Delta\bar{A}$ y $\Delta\bar{R}$ utilizados son los que se muestran en las tablas 5.2 (F_0) y 5.5 (energía).

Por otra parte, el ajuste de la duración se lleva a cabo multiplicando los valores generados por el sistema de CTH por el factor de corrección de la duración. La nueva velocidad del habla se consigue modificando la duración de las pausas y de los segmentos, teniendo en cuenta que los grupos fónicos finales de frase se tratan de forma diferente del resto de grupos fónicos.

Para evaluar el habla emocional generada automáticamente a partir de un texto, se realizó una prueba perceptiva con diez sujetos no expertos que escucharon cuatro locuciones sintetizadas a partir del mismo texto en dos fases consecutivas. El oyente tenía que escoger entre las cuatros emociones o la opción "emoción no identificada".

En la figura 5.9 se muestran los porcentajes de identificación y de confusión para las cuatro emociones en las dos pruebas. La tristeza es la emoción con mayor porcentaje de identificación, seguida por el miedo. La rabia se confunde con la alegría en un 30 %. En la primera fase, la alegría únicamente se identifica en un 20 % alcanzando el 40 % en la segunda. Este resultado permitió confirmar la hipótesis de partida, que afirmaba que la alegría era difícil de generar únicamente mediante una modificación prosódica del habla neutra.

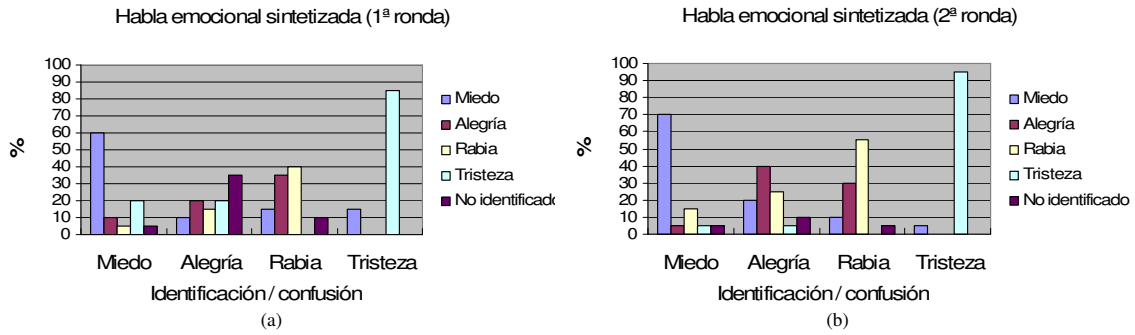


Figura 5.9: Porcentajes de identificación de la emoción obtenidos en el test de percepción realizado con muestras obtenidas tras la incorporación del módulo prosódico al sistema de CTH en catalán

5.1.3. Limitaciones de los modelos presentados y nuevo enfoque

Una vez analizados los resultados obtenidos con las dos primeras aproximaciones al modelado prosódico del habla emocional (presentadas en los apartados precedentes), se pudieron extraer conclusiones sobre la metodología seguida y los recursos utilizados. Este análisis tuvo como principal finalidad detectar las principales limitaciones que impidieron conseguir el resultado deseado y marcar los pasos hacia un nuevo enfoque que permitiese lograr una síntesis del habla expresiva de mayor calidad.

Las principales limitaciones de estas dos primeras aproximaciones a la síntesis del habla expresiva han sido las siguientes:

1. Los corpus disponibles para el modelado y la síntesis del habla no han sido suficientes para cubrir las necesidades de ambas tareas. Recordemos que el corpus de análisis inicial para el español permitió obtener unos modelos generales de la prosodia para cada una de las emociones estudiadas, pero no tenía suficiente cobertura segmental y prosódica para modelar las diferentes funciones de la prosodia y utilizarse en el proceso de síntesis. La falta de un corpus para el catalán nos obligó a generar uno de forma sintética mediante el sistema de CTH disponible en ese momento, aprovechando la similitud de los dos idiomas. El corpus para la síntesis en catalán era un corpus de dimensiones reducidas con una única instancia de cada difonema o trifonema utilizados por el sistema.

2. Uno de los aspectos más importantes que se han detectado en estos estudios previos es la dependencia temporal de la expresión oral de las emociones. Según la emoción, el habla sufre cambios en determinados parámetros, pero que a veces únicamente están presentes en ciertos segmentos del mensaje oral (Rodríguez et al., 1999; Iriondo et al., 2000). Sin embargo, la posición y la frecuencia de estos cambios no se han conseguido modelar con las aproximaciones seguidas hasta el momento. Por ejemplo, los resultados del modelado de la alegría han sido del todo insuficientes, ya que el modelo sólo se ha basado en variaciones a escala global de los parámetros prosódicos sin conseguir el efecto expresivo deseado.

3. Además, el modelado de los diferentes parámetros prosódicos tiene unas particularidades que se deben tener en cuenta para no perder determinados matices que son importantes desde el punto de vista perceptivo. Por ejemplo, el método global utilizado para el modelado de la duración no contempla que cada conjunto de fonemas y alófonos tiene su propia elasticidad (Brinckmann y Trouvain, 2003) y que no es conveniente alargar o acortar de forma uniforme todos los segmentos sintetizados. Por otra parte, el modelado de la entonación mediante la variación del tono medio y del margen dinámico a partir de un modelo para el habla neutra es insuficiente, ya que también se producen variaciones en la forma del contorno. Por último, también se han detectado problemas relacionados con el modelado de la energía, ya que el esfuerzo vocal que suponen ciertas emociones no se emula con un simple ajuste global.

4. Por último, una limitación importante ha sido la utilización de un sistema de CTH basado en la concatenación de difonemas y trifonemas de un corpus reducido y en la modificación de la señal mediante TD-PSOLA. Se considera que el habla expresiva requiere un mayor control en la modificación prosódica y en los parámetros de cualidad de la voz asociados a diferentes emociones.

Los pasos seguidos para superar las limitaciones que se acaban de detallar son los siguientes:

1. Aunque el desarrollo de un corpus de habla emocionada en cualquier idioma es una tarea difícil y costosa, se apostó por la creación de un corpus oral expresivo (véase el capítulo 4) que sirviese como núcleo de los avances en modelado prosódico y en

síntesis del habla. Como se detalla en ese capítulo, un corpus oral orientado a la síntesis del habla expresiva se debe diseñar tanto para su utilización en el modelado acústico de las emociones como para su uso en la base de datos del sintetizador de habla.

2. El hecho de disponer de un amplio corpus oral expresivo, permitiría explorar técnicas de modelado basadas en la aplicación de métodos de aprendizaje automático a estos datos. Partiendo de trabajos previos orientados a la síntesis del habla en castellano, como los presentados por Escudero (2003) sobre el modelado de la entonación y por Montero (2003) sobre síntesis del habla emocionada, se diseñó un nuevo enfoque para la predicción de la prosodia en los sistemas de síntesis del habla expresiva basado en técnicas de aprendizaje automático.

3. La mejora de los modelos obtenidos dependerá de diferentes elementos como el corpus utilizado, el análisis del texto, la definición de los atributos prosódicos, las unidades básicas de modelado de la prosodia y las técnicas de aprendizaje automático. En el caso de los tres parámetros prosódicos estudiados (F_0, energía y duración) se buscarán soluciones que permitan solventar las limitaciones detectadas adaptando los elementos del modelado a la naturaleza del parámetro.

4. Una posible solución a la mayor versatilidad del sintetizador que requiere la síntesis de habla expresiva es el uso de técnicas basadas en selección de unidades, que tienen como finalidad minimizar la modificación prosódica de la señal de voz en tiempo de síntesis (Iida et al., 2003). En consecuencia, el habla sintética resultante puede alcanzar un sonido natural para los estilos/emociones que cubre el corpus, pero a costa de utilizar corpus de gran tamaño. Las características relacionadas con la cualidad de la voz permanecen inherentes a la señal de voz en cada subcorpus asociado a un estilo o emoción determinados. Otra posible solución sería la utilización de técnicas de síntesis paramétrica del habla que permitan un mayor control para modificar la prosodia y la cualidad de la voz del habla resultante.

En este contexto se enmarca la investigación llevada a cabo para conseguir un modelado cuantitativo de la prosodia del habla expresiva utilizando técnicas de aprendizaje automático aplicadas al corpus oral descrito en el capítulo 4.

5.2. Modelado cuantitativo de la prosodia basado en corpus

En los sistemas de CTH, el modulo de predicción de los rasgos prosódicos a partir del texto de entrada es uno de los máximos responsables de la calidad del habla sintética (vid. el apartado 2.3). Además, la variación de los rasgos prosódicos está claramente relacionada con el habla expresiva (véase el apartado 2.2.1). Por lo tanto, en el ámbito de la síntesis expresiva, se deben modelar ambas funciones de la prosodia: la lingüística y la paralingüística. Dicho cometido se puede abordar mediante soluciones basadas en el conocimiento experto o recurriendo a técnicas basadas en corpus. Las primeras utilizarían un conjunto de reglas propuestas por expertos en lingüística con las que se controlaría el comportamiento de los rasgos prosódicos asociados a un texto. Desde nuestro punto de vista, este tipo de aproximación presenta los inconvenientes siguientes:

- La dificultad para representar mediante un conjunto finito de reglas la elevada variabilidad prosódica asociada a la gran diversidad del texto.
- Un conocimiento parcial del comportamiento de los rasgos prosódicos en el habla emocionada.
- La obtención de reglas normalmente cubre un objetivo muy concreto (p.ej. la función lingüística de un único estilo de habla).

Teniendo en cuenta estas dificultades y que, desde el punto de vista de la ingeniería, resulta más atractivo utilizar una aproximación basada en corpus, en la cual el conocimiento experto sobre la materia no es tan exigente, se ha optado por desarrollar un sistema de predicción de los rasgos prosódicos para el castellano utilizando técnicas de aprendizaje automático que se han aplicado al corpus oral expresivo descrito en el capítulo 4.

En este contexto, se tienen que definir: los rasgos o parámetros prosódicos del modelado, las unidades básicas para cada rasgo prosódico, los atributos prosódicos que se extraerán a partir del análisis del texto y, por último, los algoritmos de aprendizaje automático utilizados.

5.2.1. Definiciones previas

Los parámetros que determinan la prosodia de un texto oralizado son, esencialmente, la duración y la intensidad segmental, la posición y la duración de las pausas y el contorno de F_0 (Llisterri et al., 2003, , entre otros)). En el ámbito de los sistemas de CTH, la bibliografía en modelado prosódico es muy extensa, especialmente en lo que se refiere a la melodía. Sin embargo, la presencia de trabajos relacionados con la duración segmental es menor, y también es muy escasa para la intensidad (véase el resumen del apartado 2.3.1). Los parámetros que se modelarán en este trabajo son la curva de F_0, la intensidad y la duración de los sonidos del habla. Estos parámetros están relacionados perceptivamente con la entonación, el acento y el ritmo del habla.

Para cada parámetro prosódico se debe decidir cuál es la unidad acústica básica que servirá de base para su posterior modelado.

En el habla natural, la duración de los sonidos depende, entre otros factores, del contexto en el que se encuentran. La mayor parte de los estudios utilizan el fonema o el alófono como unidad básica para la duración (p. ej. Febrer et al., 1998a; Navas et al., 2002; Teixeira y Freitas, 2003), aunque existen aproximaciones basadas en unidades mayores como la sílaba (Campbell, 1990). En el presente trabajo, también se ha escogido el fonema o el alófono[3] como unidad básica para el modelado de la duración de los sonidos del habla.

La duración para cada segmento se puede representar directamente en milisegundos (ms) o en *z-score* (ecuación 4.1), una medida normalizada según la media y la desviación típica para cada segmento, estimada a partir de las duraciones de todas las instancias de cada segmento en el corpus. Las ventajas de utilizar una medida normalizada son las siguientes:

- Como indican Navas et al. (2002), la distribución de las duraciones de los sonidos del habla suele ser del tipo log-normal; mediante la transformación basada en el *z-score* se consigue acercar su distribución a la normalidad, lo que facilita su uso con determinados métodos estadísticos que requieren distribuciones normales. A modo de ejemplo, en la figura 5.10 se muestran dos histogramas con la distribución de las duraciones los fonemas y alófonos en ms y en *z-score* para el estilo neutro del corpus expresivo utilizado.
- El *z-score* se ha empleado en los sistemas de CTH para predecir las duraciones segmentales individuales y poder garantizar el principio de elasticidad. Según Schweitzer y Möbius (2003), los diferentes segmentos se pueden alargar o acortar según la propia elasticidad del segmento, y presentan un comportamiento propio que se puede modelar a partir de la media y la desviación típica. Por lo tanto, la variación de la velocidad del habla mediante el aumento o la disminución de la duración de los segmentos es más natural si se realiza a partir del *z-score* que si se lleva a cabo mediante la aplicación de un factor constante sobre la medida en ms.

La descripción de la curva de intensidad también se llevará a cabo mediante la energía de cada segmento de la frase que se desea modelar. En Blecua y Acín (1995) se presenta una propuesta de modelo de la intensidad vocálica para el castellano y el catalán aplicable a sistemas de CTH. La unidad básica escogida para la intensidad será el fonema aunque, a diferencia de Blecua y Acín (1995), se considerarán todos los segmentos y no sólo los vocálicos.

En Escudero (2003) se realiza una revisión de diversos trabajos centrados en las unidades para el modelado de la melodía y en los factores que caracterizan a cada una de

[3]Siguiendo la terminología habitual en lingüística, el fonema se concibe como una unidad distintiva de carácter abstracto e invariable, mientras que el alófono corresponde a la realización de un fonema, en general predecible y condicionada por el contexto. Empleamos el término "segmento" para hacer referencia tanto a fonemas como a alófonos.

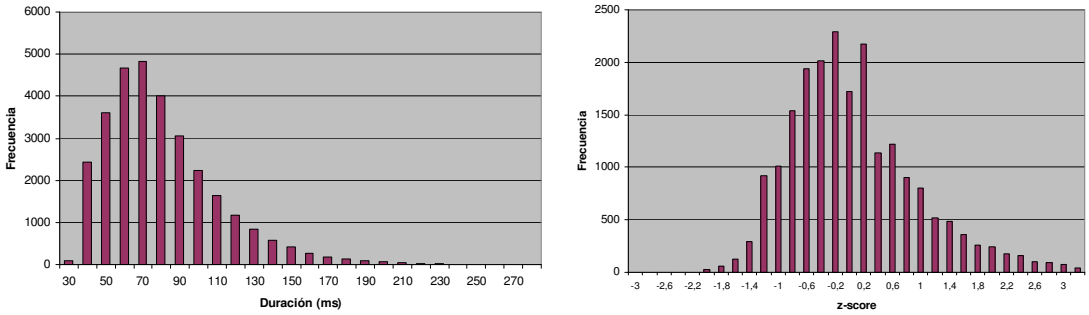


Figura 5.10: Histogramas con la distribución de las duraciones segmentales para el estilo neutro en ms y *z-score*

ellas. Para modelar el contorno melódico se ha recurrido a diferentes tipos de unidades: las unidades inferiores a la sílaba y la sílaba —relacionadas con la micromelodía—, el grupo acentual (GA) —relacionado con el ritmo del habla—, el grupo entonativo (GE) y otras unidades superiores que intervienen en la planificación del discurso. Siguiendo la propuesta de dicho autor, utilizaremos el GA como unidad básica para el modelado de la melodía. Si el lector desea profundizar en la definición de las unidades de entonación puede consultar Garrido (1996).

En este trabajo, se define el GE como una estructura coherente de entonación que no incluye ninguna ruptura prosódica importante. La separación de un texto en los correspondientes GE estará guiada por los signos de puntuación, ya que un delimitador natural de los GE es la pausa.

Entre las posibles definiciones de GA, se ha escogido la siguiente: palabra acentuada precedida, si es el caso, por una o más palabras átonas[4]. El principal motivo de escoger esta definición es que simplifica el proceso de segmentación del texto en los GA que lo componen, ya que las palabras están claramente delimitadas por espacios en blanco. En cambio, utilizar una definición basada en la sílaba en vez de la palabra implica una correcta descomposición silábica.

La curva de F_0 de cada GA se representa cuantitativamente mediante los coeficientes de un polinomio aproximador de grado n (ecuación 5.3). Para encontrar los coeficientes del polinomio, se parte de una colección de puntos $(t_i, F_0(t_i))$ que representan el valor de la F_0 media de los segmentos que forman el GA. Este valor de F_0 media se refiere al instante central del segmento. Mediante el método de mínimos cuadrados se calcula el polinomio aproximador de grado n que minimiza el error dado en la fórmula 5.4.

$$\widehat{F}_0(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n \tag{5.3}$$

[4]En Garrido (2001), a esta secuencia de palabras se la denomina grupo tónico (GT) y, en cambio, el GA se define como "la secuencia formada por una sílaba tónica (con acento primario) y todas las sílabas átonas (sin acento primario) que la siguen hasta la siguiente sílaba tónica".

$$E = \sum_{i=0}^{m} \left(F_0(t_i) - \widehat{F}_0(t_i) \right)^2 \tag{5.4}$$

Como se puede observar en la parte superior de la figura 5.11, si se tratan los GA de forma independiente, se producen discontinuidades de la curva de F_0 en los puntos de transición de un GA al siguiente. Por este motivo se ha incluido información contextual en el cálculo de los coeficientes del polinomio, es decir, se tienen en consideración los puntos correspondientes a los dos segmentos adyacentes de los GA vecinos (véase la parte central de la figura 5.11). Si se trata de un GA inicial, se tiene en cuenta el valor de F_0 del primer segmento del GA siguiente. De forma análoga, se incluye en el cálculo el último fonema del GA anterior si se trata de un GA final. Además, para conseguir unificar la longitud de todos los GA, el eje temporal se normaliza entre 0 y 1, de forma que el instante 0 representa el inicio del primer segmento del GA y el instante 1, el final del último segmento del GA (parte inferior de la figura 5.11).

5.2.2. Atributos prosódicos

La extracción de los atributos prosódicos necesarios para la predicción de la duración segmental y de los contornos de F_0 y energía a partir del texto se realiza de forma automática mediante la herramienta de análisis lingüístico descrita en el anexo E.1. Dicho *software* proporciona la transcripción fonética del texto y lo divide en grupos de entonación, grupos acentuales, palabras y sílabas.

La elección del conjunto de atributos prosódicos utilizado en la implementación práctica del módulo de predicción de los rasgos prosódicos ya mencionados es fruto del estudio de la bibliografia relacionada con este tema, de la funcionalidad del *software* de análisis del texto y de una serie de pruebas preliminares realizadas con diferentes subconjuntos de atributos. Finalmente, se han utilizado los atributos mostrados en la tabla 5.6 para la duración, la intensidad y la F_0 respectivamente. En esta tabla se recogen las etiquetas empleadas, una breve descripción de cada una y el tipo de atributo. Como se describe en el apartado 5.3, también se ha estudiado la inclusión de un atributo con información morfológica para la predicción de la duración segmental y de la curva de F_0.

A fin de modelar la duración segmental y la intensidad, se ha escogido el segmento (fonema o alófono) como unidad acústica básica. Como se puede observar en la tabla 5.6, la predicción de la duración de un segmento se realizará a partir de su identidad (FON1) y del contexto donde se encuentra, representado en este caso por el segmento anterior (FON0) y el siguiente (FON2). También se ha considerado si el segmento pertenece a una sílaba tónica (ACENTUADO) y su posición en la frase que se representa mediante dos atributos: la posición del GA al que pertenece el segmento dentro de su GE (GA-en-GE) y la posición del segmento en el GA (FON-en-GA). Cada uno de estos dos últimos atributos permiten distinguir cuatro casos: inicial, central, final o único.

Para la intensidad, se han escogido atributos relacionados con la identidad del segmento (FON1), con su pertenencia a una sílaba tónica o átona (ACENTUADO) y con

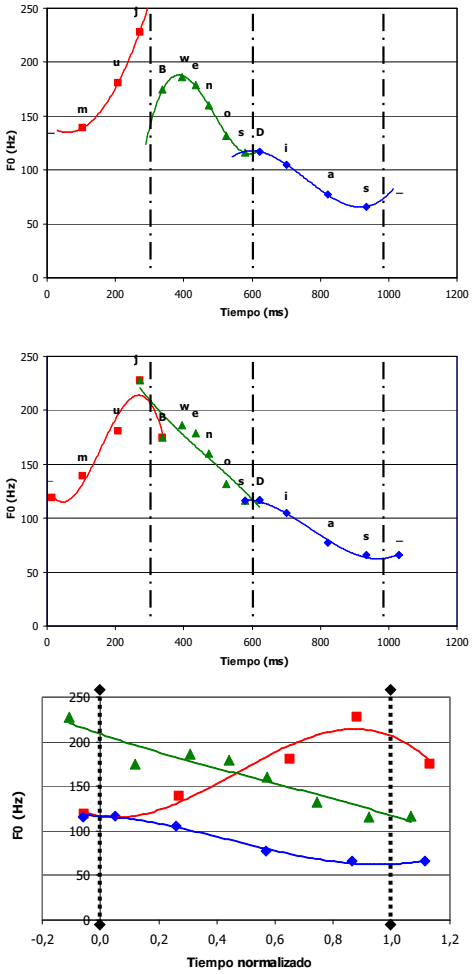


Figura 5.11: Polinomios aproximadores de los tres GA que forman el GE "Muy buenos días" sin información contextual (figura superior) y teniendo en cuenta los valores de F_0 del último segmento del GA anterior y del primer segmento del GA siguiente (figura central). En la figura inferior se muestra el proceso de normalización del eje temporal.

su posición, que se representa por tres atributos: la posición del GA en el GE (GA-en-GE) —el mismo que para la duración—, la posición numérica del segmento en el GA (FON-en-GA) y la posición en el GE (FON-en-GE), que distingue si se trata de un segmento inicial, central o final respecto a esta unidad entonativa (véase la tabla 5.6).

Para el modelado de la curva de F_0, se ha elegido el GA como unidad básica siguiendo la propuesta de Escudero (2003) (véase el apartado 5.2.1). El GA incorpora la influencia de la sílaba (cada GA está compuesto de una sílaba tónica y de las sílabas átonas precedentes) y la estructura en el nivel de GE se consigue mediante la concatenación de los GA que lo forman. Por lo contrario, este modelo no toma en consideración las variaciones debidas a la micromelodía.

La selección del conjunto de atributos utilizado para la predicción de la curva de F_0 se ha basado en propuestas realizadas para el español en las que se aborda la CTH

Tabla 5.6: Atributos prosódicos para la predicción de la duración, la energía y la F_0

Etiqueta	Atributo	Tipo *
FON0	Fonema o alófono anterior	D
FON1	Fonema o alófono actual	D
FON2	Fonema o alófono siguiente	D
ACENTUADO	Fonema o alófono acentuado	B
GA-en-GE	Posición del GA en el GE	D
FON-en-GE	Posición de FON en el GE	D
Duración	Duración del fonema o alófono en *ms*	N

Etiqueta	Atributo	Tipo
FON1	Fonema o alófono actual	D
ACENTUADO	Fonema o alófono acentuado	B
GA-en-GE	Posición del GA en el GE	D
FON-en-GA	Posición del FON en el GA	N
FON-en-GE	Posición del FON en el GE	D
Energía	Energía del fonema o alófono en *rms*	N

Etiqueta	Atributo	Tipo
TIPO-GE:	Tipo de GE	D
GA-en-GE	Posición del GA en el GE	D
ACENTO	Posición de la sílaba tónica	D
GA-en-FRA	Posición del GA en la frase	D
NUM-SIL	Número de sílabas del GA	N
$\mathbf{a}_0, \mathbf{a}_1, \ldots \mathbf{a}_n$	Coeficientes del polinomio aproximador del contorno de F_0 de un GA	A

* (D) Discreto, (B) Binario, (N) Numérico, (A) *Array* numérico

desde una perspectiva eminentemente tecnológica (Escudero et al., 2002, 2003; Campillo y Rodríguez, 2006). Por el momento, sólo se diferencia entre GE enunciativos, interrogativos, exclamativos o en suspensión[5], que se detectan fácilmente a partir de los signos de puntuación (atributo TIPO-GE). El atributo ACENTO indica la posición de la sílaba tónica en el GA, distinguiéndose cuatro posibilidades: agudo, llano, esdrújulo o sobresdújulo. La posición del GA en la frase (GA-en-frase) se ha cuantificado en cuatro valores: inicial, central, final o única en el caso de que la frase tenga un único GA. El número de sílabas (NUM-SIL) está relacionado con la longitud del GA (véase la tabla 5.6).

5.2.3. Modelado automático de la prosodia mediante CBR

El aprendizaje automático —*Machine Learning*— (ML) comprende un conjunto de técnicas que permiten reconocer una situación problemática y reaccionar utilizando la estrategia aprendida para un nuevo problema. La utilización del ML puede ser interesante en aquellos dominios en los que la experiencia es escasa y la codificación del conocimien-

[5]Esta tipología, basada en Campillo y Rodríguez (2006), recoge, por una parte, las tres modalidades oracionales clásicas y, por otra, la existencia de enunciados inacabados.

to que la describe es limitada o fragmentaria y, por lo tanto, incompleta. La predicción de los rasgos prosódicos a partir del texto es una tarea compleja en la cual intervienen elementos lingüísticos, como los fonéticos y los pragmáticos. La utilización de técnicas de ML para dicha tarea puede deparar resultados válidos dentro del ámbito de los sistemas de CTH. En general, el aprendizaje se llevará a cabo a partir de un conjunto de muestras de entrenamiento. En Duda et al. (2001) se distingue entre aprendizaje supervisado, no supervisado y por refuerzo. En el aprendizaje supervisado, el conjunto de entrenamiento dispone de las soluciones, a diferencia del aprendizaje no supervisado, que trata de obtener agrupaciones naturales de patrones de entrada. En el aprendizaje por refuerzo la información utilizada para el entrenamiento de un sistema se obtiene a partir de la respuesta de un agente externo a las acciones del propio sistema. Las técnicas de ML se pueden clasificar principalmente en: aprendizaje analógico (razonamiento basado en casos), aprendizaje inductivo (árboles de decisión), aprendizaje evolutivo (algoritmos genéticos) y aprendizaje conexionista (redes neuronales artificiales). En general, las técnicas más utilizadas en tareas de aprendizaje automático relacionados con la predicción de rasgos prosódicos son los árboles de clasificación y regresión —*Classification And Regression Trees*— (CART) y las redes neuronales artificiales —*Artificial Neural Network*— (ANN).

Para el modelado de la **duración segmental**, se han propuesto principalmente técnicas que se enmarcan en el aprendizaje inductivo, tales como CART (Möbius y van Santen, 1996; Febrer et al., 1998a; Bagshaw, 1998; Lee y Oh, 1999; Brinckmann y Trouvain, 2003; Navas et al., 2005; Mixdorff et al., 2005; Krishna y Murthy, 2005), y en el aprendizaje conexionista como las ANN (Campbell, 1990; Riedi, 1995; Córdoba et al., 1999; Teixeira y Freitas, 2003; Montero et al., 2004). Si el tamaño del corpus disponible no es suficiente para realizar este tipo de aproximaciones estadísticas, se puede llevar a cabo una regresión lineal como proponen Mixdorff et al. (2003). Según Lee y Oh (1999), el uso más extendido de CART respecto a otros métodos como las ANN se debe, en parte, a que posibilita una mejor comprensión del proceso de predicción.

Aunque en menor medida, también existen en la bibliografía aproximaciones para la predicción del **contorno de energía** basadas en las mismas técnicas de ML que para la duración: CART (Bagshaw, 1998; Lee et al., 2000) y ANN (Lee et al., 1998).

La bibliografía sobre el modelado de la melodía es la más extensa del conjunto de los tres rasgos prosódicos considerados. La generación de **contornos de $\mathbf{F_0}$** es un problema que se ha abordado mediante diferentes técnicas de ML. Por ejemplo, la utilización de ANN está presente en los trabajos de Montero et al. (2003, 2004), y una extensión de CART para la predicción de vectores que modelan el contorno de F_0 se presenta en Agüero et al. (2004), en ambos casos aplicados al español.

También existe alguna aproximación que genera más de un rasgo prosódico de forma simultánea como, por ejemplo, una red neuronal recurrente utilizada en Farrokhi et al. (2004) que permite la estimación del contorno de F_0, del contorno de energía y de la duración de sílabas, vocales y pausas para un sistema de CTH para el persa.

Los sistemas de predicción de la prosodia a partir de texto que utilizan algoritmos de ML están basados en un aprendizaje realizado sobre una base de datos o corpus (del

inglés *data-driven* o *corpus-based*). Este aprendizaje se lleva a cabo en la fase de entrenamiento que, mayoritariamente, suele ser supervisado. De las propuestas mencionadas en este apartado, únicamente Bagshaw (1998) propone un aprendizaje no supervisado. El modelado de la prosodia basado en corpus propuesto en el presente trabajo se esquematiza en la figura 5.12, que muestra un diagrama de bloques que diferencia una fase de entrenamiento supervisado y una fase de explotación, en la cual el sistema ha de predecir automáticamente la prosodia que corresponde a la oralización del texto de entrada. El nexo de unión entre las dos fases son los algoritmos de ML, una vez que los datos se han adaptado para su utilización. El proceso de entrenamiento parte de la información almacenada en el corpus de habla expresiva, y se preparan las muestras para cada uno de los tres parámetros que se deben predecir, los cuales serán procesados según el algoritmo de ML utilizado. Se trata de un entrenamiento supervisado a partir de un conjunto de muestras formadas por los atributos prosódicos (A) extraídos del análisis del corpus, más el valor (V) de la clase numérica, que es el parámetro que es necesario predecir en la fase de explotación. En esta fase la entrada es un texto y el proceso de cálculo de los atributos prosódicos asociados al texto es idéntico al del entrenamiento, a excepción de la clase numérica que, en este caso, es el valor que se pretende predecir. Finalmente, una vez estimados los parámetros prosódicos y, junto con la transcripción fonética, el módulo de síntesis de voz es el encargado de generar la versión sonora del texto. La calidad del habla sintetizada dependerá en gran medida del funcionamiento del módulo de predicción de la prosodia.

5.2.3.1. Fundamentos del CBR

El razonamiento basado en casos —*Case Based Reasoning*— (CBR) es un tipo particular de aprendizaje analógico. La analogía trata de resolver un problema objetivo a partir de la experiencia acumulada en la resolución previa de uno o más problemas base (Moreno et al., 1994). Se parte de la hipótesis que si dos situaciones o casos son similares de base en algún aspecto, también pueden serlo en algún otro.

Según Aamodt y Plaza (1994), el ciclo principal del CBR puede descomponerse en cuatro tareas (Ciclo 4R): recuperar los casos mas similares (*retrieve*), adaptarlos para resolver el problema (*reuse*), revisar la solución propuesta (*revise*) y aprender de la experiencia (*retain*), como se puede observar en la figura 5.13. El corazón del sistema es la memoria de casos resueltos, que se debe inicializar correctamente.

El principal problema puede ser el coste en memoria, ya que se suele trabajar con un alto volumen de casos cuando la complejidad es elevada. El primer paso consiste en inicializar la memoria de casos de forma que sea representativa, lo más compacta posible, y que esté bien organizada.

El objetivo de la tarea de **recuperación** es encontrar la solución desde la memoria de casos al nuevo problema. Se recupera el caso (o los casos) más similar utilizando una métrica adecuada a los atributos que lo representan. Existen diferentes funciones para comparar dos casos a partir de los atributos que los representan. En la implementación

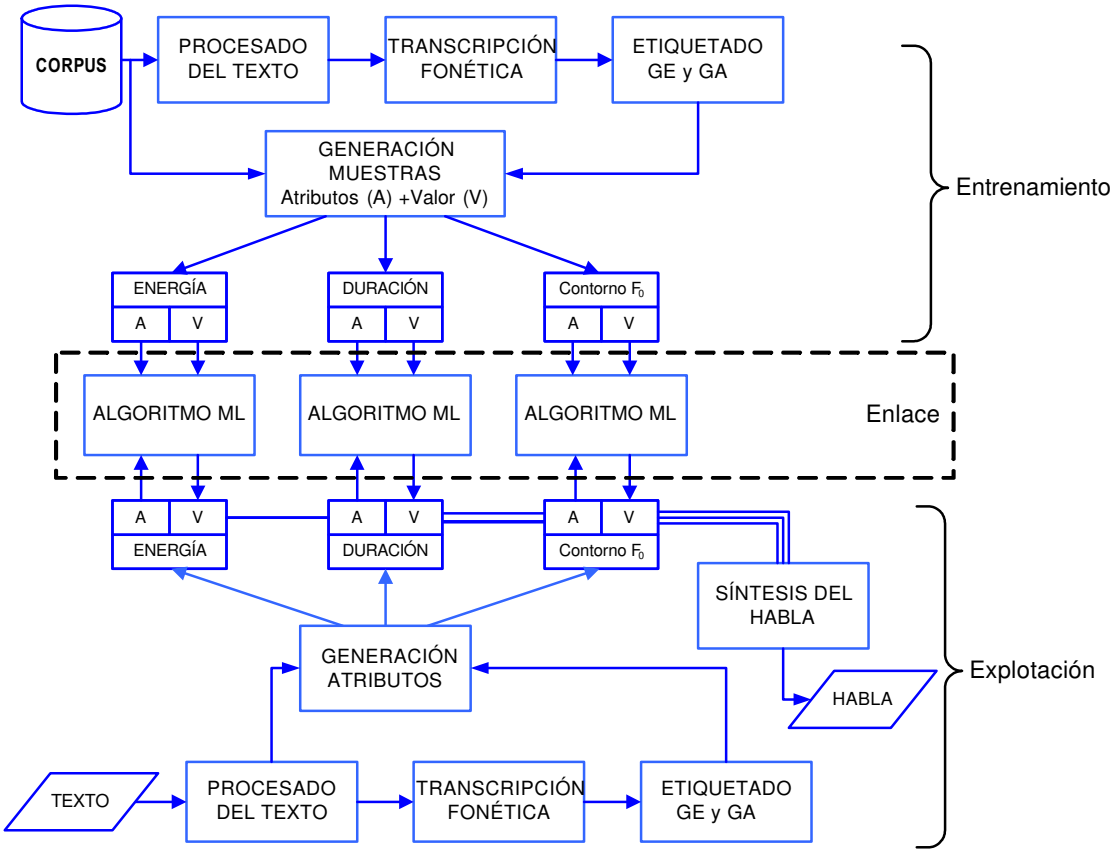


Figura 5.12: Esquema de los procesos de entrenamiento y de explotación en el modelado prosódico basado en corpus

realizada se ha utilizado la métrica de Minkowski, que viene dada por la fórmula 5.5:

$$d(x,y) = \sqrt[r]{\sum_{i=1}^{F} w_i \left|x_i - y_i\right|^r} \tag{5.5}$$

donde x e y representan los atributos de los casos que se comparan, w es un vector de pesos con el que ponderar los diferentes atributos, F es el número de atributos y, según el valor de r, se establecen tres variantes: $r = 1$ Hamming, $r = 2$ Euclidiana y $r = 3$ Cúbica.

El objetivo de la fase de **adaptación** consiste en adecuar la solución obtenida a la naturaleza del nuevo problema. Por ejemplo, en el caso de un clasificador, si se recupera más de un caso, la clasificación final se puede obtener mediante la votación mayoritaria. En el caso de la predicción de un valor numérico se pueden promediar las diferentes soluciones. Si el problema no requiere una adaptación, entonces se da por válida la solución del caso recuperado.

La fase de **revisión** tiene como objetivo evaluar la solución propuesta por parte del usuario del sistema. Esta fase se repite hasta que la solución se considere buena o bien se decida que no se puede resolver el problema.

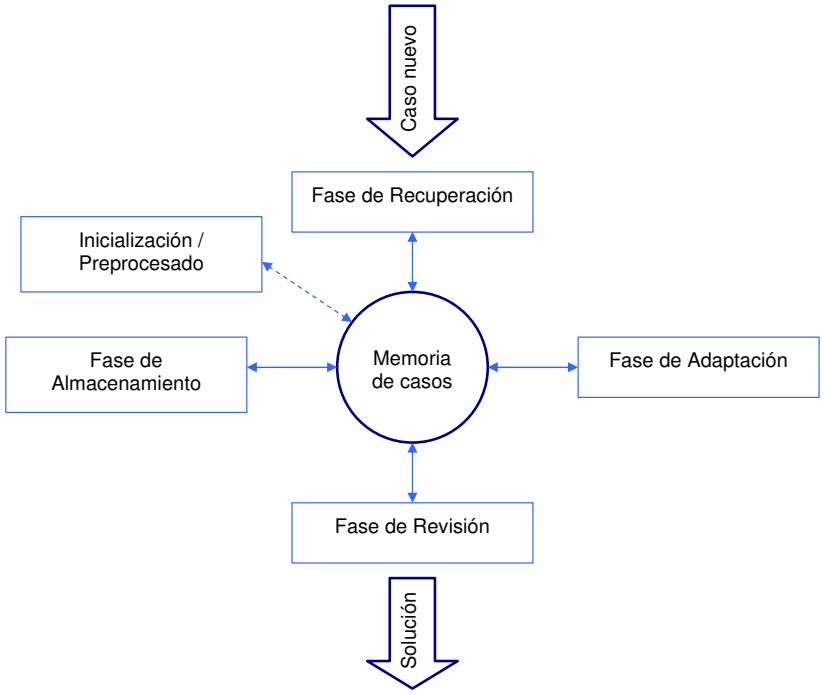


Figura 5.13: Ciclo 4R del CBR

Por último, la fase de **almacenamiento** permite incorporar conocimiento nuevo a la memoria de casos. Si la solución obtenida es buena y se trata de un caso diferente a los ya existentes, su incorporación a la memoria de casos puede mejorar el funcionamiento del sistema. En cambio, si la solución obtenida es incorrecta, será necesario reorganizar la memoria para evitar el mismo error en un futuro. Además, se puede incorporar una estrategia de olvido para ir eliminando aquellos casos que no se utilizan y así agilizar el proceso de búsqueda.

En resumen, los principales aspectos favorables del CBR son: *i)* se trata de un sistema sencillo, de fácil comprensión e implementación; *ii)* permite un tratamiento directo de los atributos que no son nominales (números reales o enteros), sin requerir la discretización de los datos; *iii)* admite una clase numérica de uno o más elementos.

Como aspectos negativos de esta técnica hay que destacar que se trata de un sistema costoso en cuanto a memoria y que, por lo tanto, una mala organización de la memoria de casos puede ralentizar el sistema. Sin embargo, el tamaño de la memoria de casos puede reducirse aplicando técnicas de *clustering*. Otra crítica habitual es que la resolución de problemas mediante CBR es opaca, ya que el usuario no puede seguir los pasos que han conducido a la solución final, a diferencia, por ejemplo, de los árboles de decisión. Sin embargo, según Grachten (2006), uno o más casos parecidos pueden dar idea de la solución al nuevo problema. Además, si la complejidad del modelo es alta, el trazado de una solución por un árbol de decisión de cientos de ramas no ayudará a aclarar la solución obtenida.

5.2.3.2. Utilización del CBR en el modelado prosódico

El modelado prosódico basado en corpus se puede abordar utilizando diferentes técnicas de ML. Se ha optado por desarrollar por completo un sistema capaz de predecir los rasgos prosódicos a partir de un texto para tener un control total de todos los pasos realizados. Esta implementación se ha llevado a cabo siguiendo la filosofía del CBR, con la ayuda de un proyecto final de carrera de Ingeniería Superior en Informática (Miralles, 2005), dirigido por el autor de esta tesis, el cual incluye un análisis de las diferentes técnicas de ML y una evaluación de las ventajas y los inconvenientes de cada una. A continuación se explica la adaptación de las diferentes fases del CBR al problema propuesto.

Como se ha comentado previamente, la **inicialización** del sistema no es propiamente una tarea del ciclo 4R del CBR, pero será imprescindible para conseguir una **memoria de casos** suficientemente representativa del dominio deseado y lo más compacta posible.

Se parte del corpus descrito en el capítulo 4, organizado en frases de las que se utilizan el texto, la transcripción fonética y los valores de duración, energía media y F_0 media de cada segmento. El objetivo del procesado lingüístico que se aplica sobre todas las frases del corpus consiste en calcular los atributos prosódicos asociados a las unidades básicas que las forman.

Siguiendo la nomenclatura propia del CBR, un caso resuelto (como los que se utilizan en el entrenamiento) está formado por un par atributos-clase, y un caso que se desea resolver (fase de explotación) está caracterizado únicamente por los atributos que permitirán encontrar la solución a partir del valor estimado de la clase.

Para la F_0, la unidad básica es el GA, que tendrá asociado unos atributos que lo caracterizarán y una clase formada por un vector numérico. Por lo tanto, el texto de entrada se deberá separar en sus GE y cada GE en los GA que lo componen. Cada GA del corpus tendrá asociados los atributos prosódicos que permitirán, una vez entrenado el sistema, predecir su curva de melodía. Para la duración segmental y la energía, la segmentación en unidades básicas es más sencilla que en el caso de la melodía, ya que directamente se trabaja en el nivel del fonema o del alófono, y el corpus utilizado ya dispone de esta información. Un ejemplo de este par atributos-clase para cada parámetro prosódico se muestra en la tabla 5.6.

El proceso de cálculo automático de los atributos prosódicos y de la clase que formarán el conjunto de casos asociados a los tres rasgos prosódicos de una frase se lleva a cabo en los pasos siguientes:

Transcripción. Mediante las funciones de la librería SINLIB (véase el anexo E.1) y las reglas desarrolladas para el castellano, el texto de cada frase se analiza con el objetivo de obtener su transcripción fonética en código SAMPA (Wells, 1993). Dichas reglas están organizadas modularmente, de forma que permiten asociar las propiedades necesarias para determinar los atributos prosódicos. Las propiedades que se establecen durante el proceso de transcripción son: el acento de la palabra, el inicio y el final de la frase junto con su tipo (determinado por los signos de puntuación),

el inicio y el final de la palabra (separadas por espacios en blanco) y la vocal que constituye el núcleo de la sílaba.

Separación en GE y GA. Partiendo de la cadena de segmentos con información sobre la acentuación, la silabificación y la delimitación en palabras y frases, asignada tal como se ha descrito en el punto anterior, se generan nuevas propiedades que son el inicio y el final del GA (según la definición establecida en el apartado 5.2.1) y el inicio y el final del GE junto con su categoría (enunciativo, interrogativo, exclamativo y en suspensión). En este punto hay que recordar que sólo se tienen en cuenta los GE delimitados por signos de puntuación y, por lo tanto, una mejora del sistema sería incorporar aquellas pausas que no vienen marcadas ortográficamente con la ayuda de un análisis morfológico y sintáctico del texto.

Inserción de la clase. En el esquema mostrado en la figura 5.12, que resume el funcionamiento global del sistema, se puede observar que en la fase de entrenamiento se insertan los valores de duración y energía de cada segmento y los parámetros de la curva de F_0 de cada GA. Los valores de energía media y de duración de cada segmento se obtienen directamente del etiquetado del corpus. Los valores de los coeficientes del polinomio aproximador que representan el contorno de F_0 de un GA se calculan siguiendo los pasos descritos en el apartado 5.2.1.

Inserción de atributos. Los atributos prosódicos descritos en el apartado 5.2.2 y resumidos en la tabla 5.6 se pueden calcular mediante unas reglas sencillas de programación, ya que simplemente se trata de reutilizar la información obtenida previamente y realizar algún cálculo simple como, por ejemplo, contar el número de sílabas de un GA o establecer la posición de una unidad dentro de otra de orden superior.

Cabe recordar que los pasos descritos son comunes para las fases de entrenamiento y explotación, con la única diferencia de la inserción de la clase (duración, energía o coeficientes del polinomio aproximador), que se da únicamente en el entrenamiento, ya que, precisamente, la fase de explotación se encarga de su estimación.

Con el objetivo de reducir el tamaño de la memoria de casos obtenida mediante el análisis de todas las frases del corpus, se agrupan todas las muestras que presentan idénticos valores para todos sus atributos. El valor de la clase resultante se obtiene a partir del promedio de las clases de los casos agrupados. Esta reducción de casos permitirá una mayor velocidad en la fase de recuperación. En la tabla 5.7 se muestra la reducción para los tres rasgos prosódicos en cada estilo del corpus expresivo. La duración y la energía presentan el mismo número de casos iniciales para cada estilo debido a que en ambas se utiliza el fonema o el alófono como unidad básica. En cambio, el número de casos para la F_0 es más reducido debido a la utilización del GA como unidad básica. La reducción de todas las memorias de casos en esta fase de inicialización es considerable, con valores comprendidos entre el 65 % y el 96 %. La reducción es mucho mayor para la energía que para la duracion, debido a que para la primera se han utilizado atributos con menos valores diferentes.

En el CBR, el entrenamiento del sistema finaliza con esta fase de inicialización, en la cual se ha obtenido la memoria de casos ya compactada.

Tabla 5.7: Reducción de las memoria de casos de duración, energía y F_0 para los 5 estilos del corpus.

Neutro			
	Casos iniciales	Casos finales	Reducción (%)
Duración	24.012	8.286	65,49 %
Energía	24.012	1.118	95,34 %
F_0	3.519	983	72,07 %
Alegre			
	Casos iniciales	Casos finales	Reducción (%)
Duración	25.267	8.490	66,40 %
Energía	25.267	1.759	93,04 %
F_0	3.548	854	75,93 %
Sensual			
	Casos iniciales	Casos finales	Reducción (%)
Duración	18.778	6.863	63,45 %
Energía	18.778	1.608	91,44 %
F_0	2.770	665	75,99 %
Agresivo			
	Casos iniciales	Casos finales	Reducción (%)
Duración	39.864	10.532	73,58 %
Energía	39.864	1.788	95,51 %
F_0	5.880	1.067	81,85 %
Triste			
	Casos iniciales	Casos finales	Reducción (%)
Duración	32.477	9.711	70,10 %
Energía	32.477	1.741	94,64 %
F_0	4.741	912	80,76 %

La explotación del sistema, es decir la predicción de los rasgos prosódicos asociados a una unidad básica a partir de los atributos prosódicos extraídos mediante el análisis del texto, consta de dos fases: recuperación y adaptación. Aunque el ciclo clásico del CBR se completa con las fases de revisión y almacenamiento, en esta implementación no se han utilizado. Por lo tanto, no se permite añadir casos nuevos en fase de explotación y, por ello, el almacenamiento se ha completado en la inicialización. Un posible uso futuro de estas dos fases consistiría en sofisticar la fase de entrenamiento incorporando un método que resolviese casos conocidos con una parte de los datos y que, mediante un proceso de evaluación objetiva o subjetiva, eliminase los casos que no dieran lugar a buenos resultados.

La fase de **recuperación** selecciona de la memoria de casos el caso (o los K casos) más similar que minimiza una medida de distancia entre los atributos del caso que se debe resolver y los atributos de los casos almacenados. La métrica utilizada viene dada por la ecuación 5.5. El caso que se debe resolver es la predicción de la duración o de la energía de los segmentos y los coeficientes del polinomio que aproxima el contorno de F_0. La distancia entre atributos numéricos (p.e el número de sílabas de un GA) presenta un cálculo directo. Sin embargo, para los atributos discretos se debe definir la distancia entre

los posibles valores que pueden tomar. En el caso de los atributos relacionados con la identidad del fonema, se ha considerado una distancia binaria que es cero en el caso de coincidir la identidad y uno en caso contrario. En cuanto al resto de atributos, éstos se han sustituido por números naturales ordenados según el parecido de los valores del atributo.

La fase de **adaptación** trata de solventar un nuevo caso a partir de la información recuperada de la memoria de casos. En primer lugar, se predice la duración de los segmentos, ya que la recuperación de la curva de F_0 se realiza sobre un eje temporal normalizado (véase la figura 5.11). La curva de F_0 se obtiene a partir de los coeficientes del polinomio recuperados. El eje temporal está normalizado entre 0 y 1 para todos los GA. Una vez conocida la duración de cada segmento se expande el eje temporal y se asocia el valor de F_0 de cada segmento según el polinomio recuperado. Por lo tanto, la duración total de cada GA es la suma de las duraciones de los segmentos que lo componen. La predicción de la energía, al igual que la de la duración, se realiza para cada fonema o cada alófono, por lo que su resultado no depende de la predicción de ningún otro parámetro prosódico.

Si se recupera más de un caso, la solución final se puede obtener o bien promediando las soluciones obtenidas o bien añadiendo un nuevo módulo de selección del mejor caso en función de los valores recuperados para las unidades vecinas. Por ejemplo, para el caso de la melodía, se podría solventar este proceso de decisión mediante la búsqueda de un camino óptimo que minimizase la discontinuidad de F_0 entre GA consecutivos, siguiendo una filosofía similar a la de la síntesis del habla basada en selección de unidades, que incorpora una función de coste de concatenación. En el presente trabajo, el valor final se obtiene mediante el promedio de los valores recuperados.

5.3. Evaluación objetiva

El funcionamiento del sistema de predicción de la duración, la energía y el contorno de F_0 se ha evaluado mediante métricas objetivas utilizando la raíz cuadrada del error cuadrático medio (RMSE) y el coeficiente de correlación de Pearson (ρ). El RMSE mide la diferencia entre los N valores observados x_k y su correspondiente estimación y_k en términos cuadráticos según la fórmula 5.6.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k - y_k)^2} \tag{5.6}$$

El coeficiente de correlación de Pearson (ρ) (ecuación 5.10) se calcula como el cociente entre la covarianza (ecuación 5.7) de dos muestras de datos x_k y y_k y sus respectivas varianzas (ecuaciones 5.8 y 5.9), siendo $\bar{x}$ e $\bar{y}$ sus medias muestrales. Mide el grado de dependencia lineal entre las dos muestras de datos. Cuando la relación es perfectamente lineal, este coeficiente presenta el valor de 1; si el valor del coeficiente se aproxima a 0, indica que no existe relación lineal.

$$s_{xy}^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) \tag{5.7}$$

$$s_x = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2} \tag{5.8}$$

$$s_y = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2} \tag{5.9}$$

$$\rho = \frac{s_{xy}^2}{s_x s_y} \tag{5.10}$$

Para cada subcorpus de voz y para cada parámetro prosódico se ha llevado a cabo una validación cruzada con 4 bloques, formados cada uno por un 75 % de casos dedicados al entrenamiento y empleando el 25 % restante para la evaluación. De esta forma, todas las muestras forman parte una vez del conjunto de test.

5.3.1. Duración segmental

Para la evaluación objetiva del módulo de estimación de la duración segmental se han probado diferentes configuraciones del sistema con el objetivo de lograr una reducción del RMSE y un aumento del valor de ρ. En primer lugar se ha configurado la fase de recuperación del CBR para obtener una única solución ($K = 1$) de la memoria de casos, es decir, el caso más similar. Se han probado cuatro vectores diferentes de pesos aplicados

a la función distancia (ecuación 5.5), que evalúa la similitud de dos casos a partir de los atributos que los representan. En la tabla 5.8 se muestran los valores de estos 4 vectores de pesos junto con los nombres que los identifican.

Tabla 5.8: Diferentes vectores de pesos utilizados en la función distancia empleada en la fase de recuperación del CBR.

	FON1	FON2	FON0	ACENTUADO	GA-en-GE	FON-en-GE
PA1	1	1	1	1	1	1
P5A1	5	1	1	1	1	1
PSel	5,6	4,4	3,4	4,2	2,2	7
P10Sel	10	4,4	3,4	4,2	2,2	7

La primera configuración de pesos (*PA1*) pondera por igual todos los atributos. El segundo vector de pesos (*P5A1*) únicamente da mayor importancia al atributo FON1, que representa al segmento para el que se está prediciendo la duración. El vector de pesos *PSel* se ha obtenido con la ayuda de las funciones de selección de atributos de *Weka*, concretamente la función *CfsSubsetEval*, que valora simultáneamente la habilidad predictiva de cada atributo del conjunto de forma individual y el grado de redundancia entre ellos, ponderando más los conjuntos de atributos altamente correlacionados con la clase, pero con baja intercorrelación. Como algoritmo de búsqueda a través del espacio de subconjuntos de atributos se ha escogido la función *GreedyStepwise*[6]. Se ha configurado la función para que proporcione una lista de atributos ordenados según su relevancia. El proceso se ha repetido para los cinco subcorpus, y se ha obtenido un único vector de pesos promediando la posición obtenida de cada atributo para los cinco subcorpus. Finalmente, se ha modificado este vector de forma manual asignando el mayor peso al atributo FON1 (*P10Sel*).

En la figura 5.14 se comparan los resultados del RMSE y de ρ en función de las cuatro configuraciones de pesos comentadas previamente. El vector *PA1* es el que obtiene el peor resultado en todos los casos. La ponderación de atributos consigue para todos los estilos que el RMSE disminuya y ρ aumente, aunque en ambos casos muy ligeramente. Con los otros tres vectores de pesos probados no se obtienen diferencias significativas en los resultados. Los siguientes experimentos que se muestran se han realizado con el vector de pesos *P10Sel* de la tabla 5.8.

El siguiente experimento ha consistido en variar el valor de K del CBR, es decir, el número de casos que se recuperan de la memoria de casos. El valor predicho se calcula promediando los valores de duración de los K casos recuperados. En la figura 5.15 se presentan los resultados de realizar un barrido con tres valores de K para cada estilo. En este experimento sí que se observa una mejora significativa de los resultados con el aumento de K. La reducción del RMSE al pasar de $K = 1$ a $K = 3$ va desde 1,7 ms (estilo

[6]Este método de selección de atributos comienza con un conjunto vacío de atributos y va añadiendo el más significativo, terminando cuando al añadir un atributo disminuye la métrica de evaluación (*forward*). También permite comenzar con el conjunto completo de atributos e ir eliminando el menos significativo (*backward*). Se ha utilizado en un modo alternativo que ordena los atributos atravesando el espacio de atributos desde el conjunto vacío hasta completarlo memorizando el orden en que los atributos se han seleccionado (Witten y Frank, 2005).

neutro) a 4,8 ms (estilo alegre). Si pasamos de $K = 1$ a $K = 5$, el RMSE disminuye entre 1,8 ms y 5,6 ms. El coeficiente de correlación presenta un aumento entre 0,03 y 0,05.

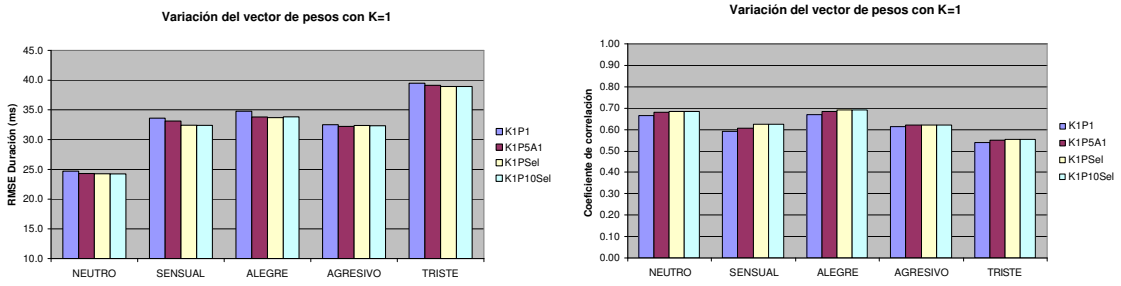


Figura 5.14: Valores de RMSE y coeficiente de correlación para la duración por estilo con el valor de K fijado a 1 y 4 vectores de pesos diferentes mostrados en la tabla 5.8

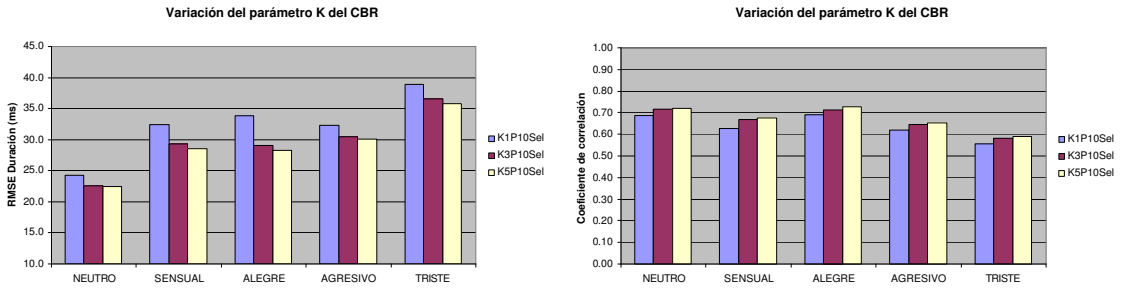


Figura 5.15: Valores de RMSE y coeficiente de correlación para la duración por estilo con $K = 1$, $K = 3$ y $K = 5$

Pueden encontrarse en la bibliografía trabajos sobre el modelado de la duración segmental que incorporan información morfológica con la intención de mejorar la predicción (p.ej. Brinckmann y Trouvain (2003)). Sin embargo, en Lee y Oh (1999), los resultados no mejoran con la inclusión de un atributo POS (del inglés *Part-of-Speech*). En el presente trabajo, no se sabe, a priori, si al añadir este tipo de información, los resultados mejorarán. Con la finalidad de estudiar este efecto, se ha añadido un atributo que incorpora información morfológica de la palabra a la cual pertenece el segmento. Las etiquetas POS se han asignado con la herramienta Freeling[7] (versión 5.1)[8], considerando únicamente el primer nivel de etiquetado, en el que se distinguen 9 categorías léxicas: adverbio, adjetivo, nombre, verbo, preposición, determinante, pronombre, conjunción e interjección.

De los cinco estilos evaluados, la alegría es el único que presenta un resultado significativamente mejor, con una reducción del RMSE de 1,8 ms. Para el estilo sensual, el RMSE sólo se reduce en 0,2 ms y, para el resto de estilos, o permanece invariable o incluso

[7]Freeling (Carreras et al., 2004) es una herramienta de análisis lingüístico desarrollada en el centro de investigación TALP de la *Universitat Politècnica de Catalunya*. El *Centre de Llenguatge i Computació* de la *Universitat de Barcelona* participó en el desarrollo de los diccionarios morfológicos y las gramáticas para el español (Atserias et al., 1998) y el catalán.

[8]`http://garraf.epsevg.upc.es/freeling/`

aumenta ligeramente (véase la figura 5.16). A la vista de estos resultados, la incorporación de este tipo de análisis no es necesaria para la estimación de la duración segmental del habla expresiva.

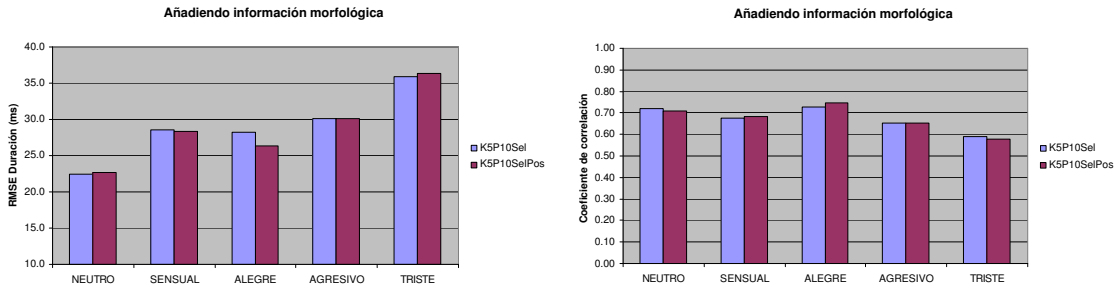


Figura 5.16: Valores de RMSE y coeficiente de correlación para la duración por estilo con y sin información morfológica. *K5P10Sel* indica un valor de $K = 5$ y el conjunto de pesos de la función distancia *P10Sel*. En la prueba *K5P10SelPos*, se añade un atributo POS.

Todos los valores medios de RMSE y de ρ que se muestran en las figuras 5.14, 5.15 y 5.16 están reproducidos en la tabla 5.9.

Tabla 5.9: RMSE medio en *ms* (a) y coeficiente de correlación medio (b) por estilo para diferentes configuraciones del sistema de predicción de la duración segmental basado en CBR

Estilo	K1P1	K1P5A1	K1PSel	K1P10Sel	K3P10Sel	K5P10Sel	K5P10SelPos
NEUTRO	24,7	24,3	24,3	24,3	22,6	22,4	22,6
SENSUAL	33,6	33,1	32,4	32,4	29,3	28,6	28,3
ALEGRE	34,8	33,8	33,7	33,8	29,0	28,2	26,4
AGRESIVO	32,5	32,2	32,4	32,3	30,5	30,2	30,1
TRISTE	39,5	39,1	38,9	38,9	36,6	35,9	36,4

(a)

Estilo	K1P1	K1P5A1	K1PSel	K1P10Sel	K3P10Sel	K5P10Sel	K5P10SelPos
NEUTRO	0,67	0,68	0,68	0,69	0,72	0,72	0,71
SENSUAL	0,59	0,61	0,63	0,63	0,67	0,68	0,68
ALEGRE	0,67	0,68	0,69	0,69	0,71	0,73	0,75
AGRESIVO	0,61	0,62	0,62	0,62	0,65	0,65	0,65
TRISTE	0,54	0,55	0,55	0,55	0,58	0,59	0,58

(b)

En resumen, de este experimento se puede concluir que de los diferentes grados de libertad del CBR, el más importante es el número de casos que es necesario recuperar. Para la tarea realizada, los mejores resultados se han obtenido con el valor de $K = 5$. Otro de los factores estudiados, el vector de pesos que pondera los diferentes atributos en la función distancia, ha mostrado que es mejor ponderar los atributos más relevantes aunque las variaciones en los resultados no sean muy grandes. Por último, se ha incorporado un atributo con información morfológica que únicamente ha proporcionado una mejora significativa para el estilo alegre. No ha aportado ningún cambio considerable para los demás estilos, e incluso ha supuesto peores resultados para los estilos triste y neutro.

Los experimentos presentados han utilizado una versión del CBR desarrollada completamente en el ámbito de esta investigación. Para validar su correcto funcionamiento se

ha considerado la posibilidad de comparar los resultados con los que pueda ofrecer una herramienta de aprendizaje automático como es *Weka* (Witten y Frank, 2005). Para cada estilo se ha adaptado la base de datos que contiene todos los pares atributos-clase al formato propio de *Weka* y se ha realizado un experimento de predicción de la duración segmental utilizando los tres métodos de regresión siguientes:

- **Regresión lineal:** Expresa la clase como una combinación lineal de los atributos a los que se les añade un atributo adicional cuyo valor es siempre 1. El entrenamiento permite calcular el valor de los pesos que multiplican los atributos mediante la minimización de la suma de las diferencias al cuadrado entre las clases reales y las predichas.

- ***Ibk:*** Se trata de regresores basados en ejemplos (*Instance Based*), que almacenan directamente las muestras de entrenamiento etiquetadas. Para predecir una nueva muestra se emplea una función de distancia para evaluar qué muestra o muestras del conjunto de entrenamiento son las más próximas a ella. En el caso del algoritmo IBk se observan las clases de los k vecinos más próximos y la predicción final se decide promediando los valores de dichas clases. Se ha probado con $K = 3$ (IBk3) y $K = 5$ (IBk5).

- ***M5P:*** Versión mejorada por Wang y Witten (1997) del algoritmo M5 (Quinlan, 1992, citado por Witten y Frank, 2005, p. 253), que implementa un árbol de modelos[9].

Los resultados de RMSE y de ρ medios obtenidos con estos cuatro algoritmos se muestran en las tablas 5.10 y 5.11, respectivamente. Se observa que para cada estilo, con y sin información morfológica, el valor más bajo del RMSE se obtiene con la versión propia del CBR. Si comparamos el mejor resultado de *Weka* (4 algoritmos y dos configuraciones de atributos) con el mejor del CBR (dos configuraciones de atributos), se observa que con el CBR se reduce el valor del RMSE un margen comprendido entre 0,4 ms (para el estilo neutro) y 2,9 ms (para el estilo triste) de forma absoluta. De forma relativa, el margen de variación va desde el 1.6 % en el estilo agresivo hasta el 7.4 % del estilo triste (véase la figura 5.17). En cuanto al coeficiente de correlación, la ventaja de un sistema sobre el otro o viceversa es mínima, destacando únicamente la diferencia del estilo triste a favor del CBR.

[9]Un árbol de modelos es un caso particular de los árboles de regresión. En un árbol de regresión, sus hojas predicen una cantidad numérica calculada como la media del valor para la variable clase de todos los ejemplos que han llegado a esa hoja durante el proceso de construcción del árbol. En cambio, las hojas de un árbol de modelos contienen una ecuación de regresión lineal local a esa partición del espacio de atributos.

Tabla 5.10: RMSE medio de la duración en ms por estilo para diferentes algoritmos de *Weka* comparado con el CBR propio.

Estilo	LR	IBk3	IBk5	M5P	CBR
NEU (POS)	24,3	23,5	23,7	23,0	**22,7**
NEU (No POS)	24,4	23,3	23,4	22,8	**22,4**
SEN (POS)	30,0	29,7	29,7	28,9	**28,3**
SEN (No POS)	30,2	29,5	29,7	28,8	**28,6**
ALE (POS)	28,5	27,5	27,4	27,1	**26,4**
ALE (No POS)	28,5	27,4	27,2	26,9	**26,2**
AGR (POS)	32,2	30,9	30,8	30,7	**30,1**
AGR (No POS)	32,4	31,0	30,9	30,7	**30,1**
TRI (POS)	39,8	39,6	39,0	39,0	**36,5**
TRI (No POS)	39,9	39,3	39,0	38,7	**35,8**

Tabla 5.11: Coeficiente de correlación medio de la duración por estilo para diferentes algoritmos de *Weka* comparado con el CBR propio.

Estilo	LR	IBk3	IBk5	M5P	CBR
NEU (POS)	0,67	0,70	0,69	**0,72**	0,71
NEU (No POS)	0,67	0,71	0,70	**0,72**	**0,72**
SEN (POS)	0,65	0,67	0,66	**0,69**	0,68
SEN (No POS)	0,65	0,67	0,67	**0,69**	0,68
ALE (POS)	0,70	0,73	0,73	0,73	**0,75**
ALE (No POS)	0,70	0,73	0,73	0,74	**0,75**
AGR (POS)	0,60	0,64	0,65	**0,65**	**0,65**
AGR (No POS)	0,60	0,64	0,64	**0,65**	**0,65**
TRI (POS)	0,53	0,54	0,55	0,56	**0,58**
TRI (No POS)	0,52	0,55	0,55	0,56	**0,59**

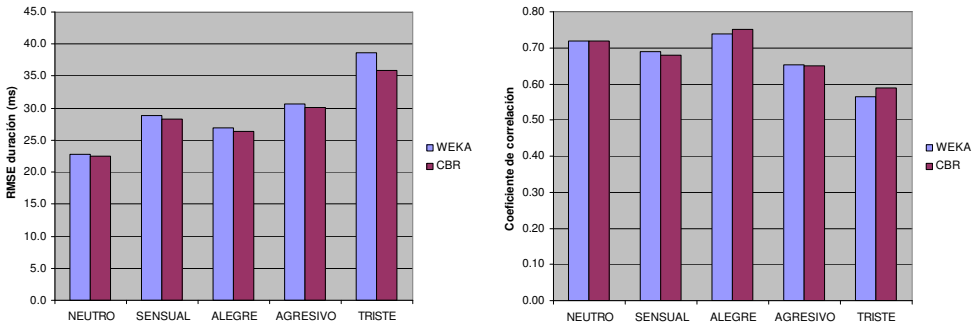


Figura 5.17: Comparación entre los mejores resultados de RMSE (izquierda) y del coeficiente de correlación (derecha) para la duración por estilo obtenidos con Weka y el CBR propio.

La comparación de resultados con otros trabajos similares se hace difícil, debido a la variedad de algoritmos de aprendizaje empleados y a las diferencias en los corpus y en los porcentajes de entrenamiento y de test. En el modelado de la duración para sistemas de CTH de dominio general no expresivos, los resultados de otros investigadores son comparables a los obtenidos para el estilo neutro (véase la tabla 5.12). Los métodos

más utilizados son ANN y CART. El mejor resultado de los mostrados (Montero et al., 2004) se corresponde a un sistema de CTH de dominio restringido y, por lo tanto, la variabilidad temporal del corpus debe de ser mucho menor que en otros corpus orientados a un dominio general.

Tabla 5.12: Resultados de diferentes estudios de modelado de la duración.

Autor/es	Idioma	Algoritmo	RMSE (ms)	ρ
Brinckmann y Trouvain (2003)	Alemán (Voz masculina)	CART	22,46	0,86
Brinckmann y Trouvain (2003)	Alemán (Voz femenina)	CART	21,40	0,83
Teixeira y Freitas (2003)	Portugués europeo	ANN	19,85	0,83
Navas et al. (2005)	Euskera	CART	22,23	0,70
Montero et al. (2004)	Castellano (Dominio restringido)	ANN	15,50	0,89
Krishna y Murthy (2005)	Hindi	CART	27,14	0,75
Krishna y Murthy (2005)	Telugu	CART	22,86	0,80

Cabe destacar que, para que una comparación entre diferentes sistemas fuera adecuada, se debería utilizar el mismo corpus. Si se da el caso de que los datos son diferentes, la utilización de una medida relativa del error puede compensar la dificultad de modelar conjuntos de datos con variabilidad distinta (Córdoba et al., 2002). El RMSE relativo tiene en cuenta la varianza de los datos ya que el error es relativo al que se cometería utilizando una simple predicción realizada con el promedio de los datos de entrenamiento (Witten y Frank, 2005). En la tabla 5.13 se muestran los valores de RMSE relativo para la predicción de la duración para cada estilo, en la que se observa que el error relativo menor se consigue para el estilo alegre, mientras que el mayor se produce con el estilo triste.

Tabla 5.13: RMSE relativo de la duración por estilo con CBR

Estilo	RMSE relativo
NEU	0.70
SEN	0.73
ALE	0.68
AGR	0.77
TRI	0.82

5.3.2. Melodía

La evaluación objetiva del módulo de estimación de la F_0 se ha basado también en las medidas de RMSE y de ρ calculadas para cada frase de test a partir del valor medio de F_0 asociado a cada segmento. De esta forma, para cada frase se obtiene un valor de estas medidas calculándolas con los valores predichos y con los pertenecientes a la misma frase del corpus. Los valores obtenidos de RMSE y de ρ para cada bloque de validación cruzada son una media de los valores obtenidos para las frases que lo forman, ponderada según el número de GA que contienen.

Los diferentes resultados que se muestran a continuación pretenden analizar el funcionamiento del sistema según los parámetros que son configurables. Una búsqueda

exhaustiva de todas las combinaciones de valores de los parámetros sería excesivamente costosa en tiempo y en computación y, por lo tanto, se ha seguido una metodología paso a paso, en la que mediante aproximaciones sucesivas, se modifican los parámetros pertenecientes a un mismo elemento del sistema y se fijan según el mejor resultado obtenido; con esta configuración se estudia el comportamiento modificando otros parámetros, y así sucesivamente. Los elementos que se han analizado mediante este método heurístico se refieren a los parámetros del algoritmo CBR, al modelo de contorno de F_0 y, finalmente, al conjunto de atributos prosódicos.

En primer lugar se ha estudiado el comportamiento del sistema CBR con diferentes vectores de pesos en la función distancia (ecuación 5.5). Para este análisis, se ha fijado a 3 el grado de los polinomios (ecuación 5.3) que aproximan el contorno de F_0 de los GA, y la recuperación del caso más parecido ($K = 1$). En la tabla 5.14 se muestran los valores de los 3 vectores de pesos utilizados junto con el nombre que los identifica.

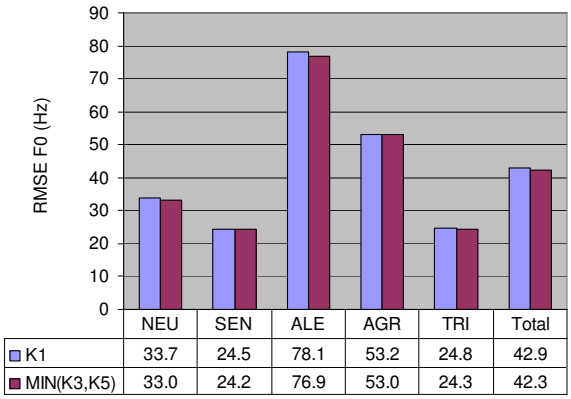
Tabla 5.14: Diferentes vectores de pesos de la función distancia utilizada en la fase de recuperación del CBR para la estimación de F_0.

	TIP0-GE	GA-en-GE	ACENTO	GE-en-FRA	NUM-SIL
PA1	1	1	1	1	1
PSel1	4,6	4,4	2,0	2,6	1,4
PSel2	3,5	4,1	2,3	2,4	2,8

La primera configuración de pesos (*PA1*) pondera todos los atributos por igual. A diferencia de la duración, ningún atributo puede considerarse a priori más relevante que los otros. Por lo tanto, los vectores de pesos *PSel1* y *PSel2* se han obtenido con la ayuda de las funciones de selección de atributos de *Weka*, concretamente con la función *CfsSubsetEval* y con el algoritmo de búsqueda *GreedyStepwise* (véase el apartado 5.3.1). La llamada a esta función de selección de atributos únicamente se puede realizar con conjuntos de datos que tengan una única clase. Debido a que para la F_0 la clase está formada por los coeficientes del polinomio, esta función sólo se puede llamar de forma independiente para cada elemento de la clase. Se han contemplado dos estrategias para estudiar la relevancia de los atributos y, de esta manera, poder ajustar el vector de pesos. En primer lugar, sólo se tiene en cuenta el coeficiente del término independiente del polinomio, que es el que está relacionado con la F_0 media del GA (vector de pesos *PSel1*). En segundo lugar, se ha obtenido un vector de pesos (*PSel2*) promediando la posición de cada atributo según el orden de relevancia obtenido al ejecutar la función de selección de atributos para cada coeficiente de forma independiente. La ponderación de pesos realizada con ambos vectores no ha supuesto una variación significativa de los resultados, observándose una ligera disminución del RMSE ($< 1\,\%$).

El otro parámetro importante del CBR es el número de casos que se recuperan (K), al que se han asignado los valores de 1, 3 y 5. Al igual que sucede con la duración, los resultados mejoran cuando se recupera más de un caso. El RMSE menor se obtiene para $K = 3$ o $K = 5$, en función del resto de parámetros y del estilo. La figura 5.18 ilustra este resultado.

Respecto a la parametrización del contorno de F_0 mediante polinomios aproxima-



	NEU	SEN	ALE	AGR	TRI	Total
K1	33.7	24.5	78.1	53.2	24.8	42.9
MIN(K3,K5)	33.0	24.2	76.9	53.0	24.3	42.3

Figura 5.18: Valores de RMSE y de ρ para la F_0 por estilo obtenidos con diferentes valores de K del CBR.

dores para cada GA, se puede variar el grado del polinomio y la continuidad entre GA vecinos. Las pruebas realizadas muestran que la predicción mejora en todos los casos si se utilizan los valores de F_0 de los segmentos adyacentes al primero y al último segmento de cada GA para aproximar los polinomios. Respecto al grado del polinomio, los mejores resultados se obtienen con polinomios de cuarto grado ($G = 4$).

Por último, se ha completado el conjunto de atributos con un atributo que incorpora la categoría morfológica (POS) de la palabra tónica del GA. Los resultados mejoran para los estilos alegre y agresivo, que son los que presentan un rango de valores de F_0 más amplio. Para el resto de estilos, los resultados son prácticamente los mismos.

En la figura 5.19 se muestran los resultados del RMSE y de ρ para cada estilo y el promedio de un subconjunto representativo de las diferentes configuraciones que se han probado. La mejor configuración se obtiene con los parámetros $K = 5$ y $G = 4$ con o sin atributo POS según el estilo. Si nos fijamos en el RMSE (figura de la izquierda), el mayor error se produce en los estilos alegre y agresivo, que son los que presentan variaciones importantes en el contorno de F_0 (véase la tabla 5.15). En los estilos sensual y triste el error es más pequeño debido a que se trata de los estilos con menores variaciones de F_0. Sin embargo, los valores más altos de ρ se obtienen con los estilos agresivo, neutro y alegre, mientras que los valores más bajos corresponden al estilo triste y al sensual. En la tabla 5.15 también se muestran las medias y las desviaciones estándares de F_0 para los GA en cada estilo. Se constata una relación directa entre el RMSE y la desviación estándar.

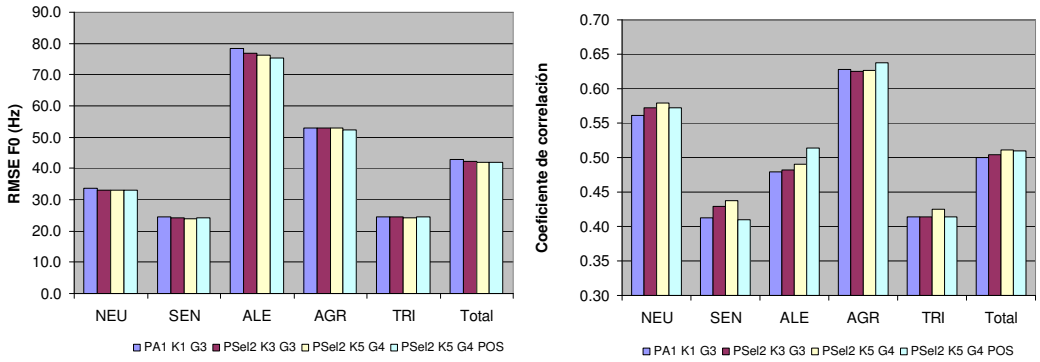


Figura 5.19: Valores de RMSE y de ρ para la F_0 por estilo obtenidos con diferentes configuraciones del CBR.

Tabla 5.15: Valores de RMSE, de ρ y de RMSE relativo para F_0 por estilo obtenidos con las mejores configuraciones individuales del CBR, junto con la media y la desviación estándar de F_0.

	RMSE (Hz)	ρ	RMSE relativo	Media	Desviación estándar
NEU	32,96	0,58	0.83	167,39	40,98
SEN	23,88	0,44	0.93	134,06	26,05
ALE	75,34	0,52	0.87	270,97	89,06
AGR	52,41	0,64	0.79	263,90	68,46
TRI	24,33	0,43	0.95	176,16	26,87

Al igual que para la duración, se ha comparado el funcionamiento del CBR propio con diferentes algoritmos regresores disponibles en *Weka*. La utilización de *Weka* no ha sido tan directa como en el caso de la duración, ya que para la F_0 la clase que se debe predecir es el vector de coeficientes del polinomio que aproxima el contorno de F_0 para cada GA. Se han desarrollado una serie de funciones en Matlab que han permitido utilizar las funciones de *Weka* mediante llamadas en Java. Para cada algoritmo de clasificación se ha llevado a cabo un proceso de validación cruzada de cuatro bloques (75 % de casos para el entrenamiento y 25 % para el test) con el fin de estimar de forma independiente cada coeficiente del polinomio. Finalmente, para cada caso del test (un GA) se han combinado la predicciones individuales para calcular los valores de F_0 y, de esta forma, poder medir el RMSE respecto al mismo caso presente en el corpus. Se han tomado 11 valores de la variable independiente (tiempo normalizado entre 0 y 1; véase el apartado 5.2.1) para ambos vectores de coeficientes: el real y el estimado. Para poder comparar el CBR se han obtenido los valores de los coeficientes del polinomio de cada GA en la fase de recuperación sin la necesidad de desnormalizar el eje temporal.

Los algoritmos de *Weka* utilizados son los mismos que para la estimación de la duración (véase la descripción en el apartado 5.3.1). El CBR utilizado está configurado con los siguientes valores: conjunto de pesos *PSel2*, $K = 5$, $G = 4$. Los resultados obtenidos se muestran en la tabla 5.16, observándose unos resultados muy parecidos del CBR con sus homólogos de *Weka* (IBK3 o IBK5). En cambio, con el algoritmo M5P se obtienen unos resultados mucho peores que con el resto de algoritmos, a diferencia de lo que sucedía

en el caso de la duración; de los algoritmos de *Weka* probados, los mejores resultados se obtuvieron con M5P. El motivo podría radicar en el hecho de tratar de predecir los coeficientes con regresores entrenados de forma independiente y juntar los resultados a posteriori. Agüero et al. (2004) han solucionado este problema utilizando V-CART, una modificación de CART adaptada a la predicción de vectores de datos. La regresión lineal y los algoritmos basados en casos no se ven afectados por este problema.

Tabla 5.16: RMSE medio de la F_0 por estilo para diferentes algoritmos de *Weka* comparado con el CBR propio configurado con los siguientes valores: conjunto de pesos *PSel2*, $K = 5$, $G = 4$ con y sin atributo POS.

Estilo	LR	IBK3	IBK5	M5P	CBR
NEU (POS)	36,33	31,49	31,21	61,71	**30,96**
NEU (No POS)	34,71	**30,41**	30,66	43,42	31,10
SEN (POS)	27,62	22,68	22,24	44,58	**22,13**
SEN (No POS)	26,69	22,15	**21,97**	31,91	22,05
ALE (POS)	86,07	68,66	68,21	184,32	**67,85**
ALE (No POS)	80,36	68,12	**67,53**	132,16	68,59
AGR (POS)	55,06	47,98	**47,52**	96,70	47,66
AGR (No POS)	51,82	47,40	**47,25**	76,30	48,23
TRI (POS)	23,61	23,36	22,94	33,17	**22,77**
TRI (No POS)	23,10	22,65	22,68	25,96	**22,56**

La comparación directa de resultados con otros trabajos de modelado cuantitativo de la F_0 para la síntesis del habla expresiva es compleja debido a la utilización de diferentes corpus y medidas de evaluación. En Tesser et al. (2005) se presentan los resultados de la predicción de F_0 mediante CART para siete emociones obtenidas de la base de datos E-Carini. El RMSE oscila entre 28 Hz obtenidos para el asco y 54 Hz para la alegría, mientras que los mismos autores habían alcanzado un RMSE de 36.5 Hz y ρ de 0.43 para la síntesis del hablea en un estilo narrativo (Tesser et al., 2004). En cambio, el RMSE disminuye significativamente en aquellos estudios en los que el corpus no es expresivo. Por ejemplo, Montero et al. (2004) consiguieron un RMSE de 19,8 Hz para un sistema CTH de dominio restringido utilizando ANN. Finalmente, con el sistema MEMOInt (Escudero y Cardeñoso, 2007) se ha obtenido un RMSE de 18,71 Hz para el conjunto de frases enunciativas del corpus utilizado en el sistema de síntesis del habla por concatenación de unidades del TALP[10].

5.3.3. Energía

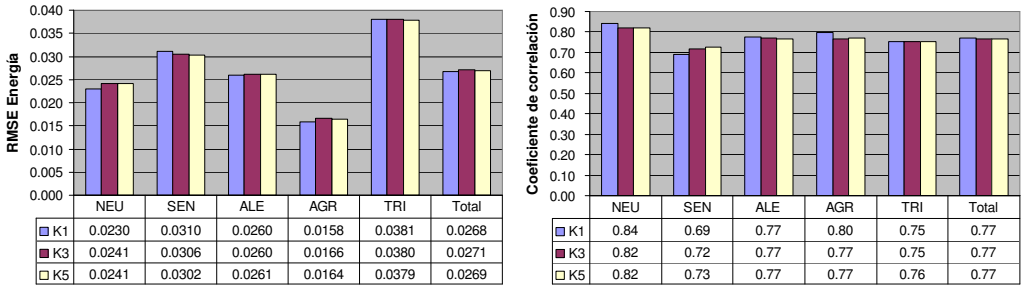
La evaluación objetiva del módulo de estimación de la energía se ha basado también en las medidas de RMSE y coeficiente de correlación (ρ) calculadas para cada frase de test a partir del valor medio de energía *rms* asociado a cada segmento. De esta forma, para cada frase se obtiene un valor de estas medidas calculándolas con los valores estimados y con los pertenecientes a la misma frase del corpus. La profundidad del estudio de la energía ha sido menor que el realizado para la duración y para la curva de F_0, ya que

[10]Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla. `http://www.talp.upc.edu/talp`.

el objetivo principal consistía simplemente en disponer de un sistema de predicción de energía aprovechando el desarrollo realizado.

Para la energía, al igual que la duración, la unidad básica escogida es el fonema o el alófono. Los resultados obtenidos para diferentes valores de K se muestran en la figura 5.20. A la vista de los resultados, no se observa una tendencia común para todos los estilos respecto el valor de K. Se obtienen mejores resultados con $K = 1$ en los estilos neutro, alegre y agresivo y, en cambio, en los estilos sensual y triste, tanto el RMSE como ρ son mejores para $K = 5$. En los tres casos, el atributo con mayor ponderación ha sido la identidad del segmento.

La energía es el parámetro menos estudiado en la investigación relacionada con el modelado prosódico orientado a la síntesis del habla. En el presente trabajo se ha aprovechado la arquitectura definida para la predicción de la duración de los segmentos adaptándola a la predicción de la energía, obteniéndose unos resultados satisfactorios desde el punto de vista de las medidas objetivas efectuadas. La utilización de diferentes algoritmos de *Weka* conduce prácticamente a los mismos resultados que con el CBR (véase la tabla 5.17). En la tabla 5.18 se muestran los valores del RMSE relativo.



	NEU	SEN	ALE	AGR	TRI	Total
K1	0.0230	0.0310	0.0260	0.0158	0.0381	0.0268
K3	0.0241	0.0306	0.0260	0.0166	0.0380	0.0271
K5	0.0241	0.0302	0.0261	0.0164	0.0379	0.0269

	NEU	SEN	ALE	AGR	TRI	Total
K1	0.84	0.69	0.77	0.80	0.75	0.77
K3	0.82	0.72	0.77	0.77	0.75	0.77
K5	0.82	0.73	0.77	0.77	0.76	0.77

Figura 5.20: Valores de RMSE y de ρ para la energía por estilo con $K = 1$, $K = 3$ y $K = 5$.

Tabla 5.17: RMSE medio de la energía por estilo para diferentes algoritmos de Weka comparado con el CBR propio.

Estilo	LR	IBK3	IBK5	M5P	CBR
NEU	0,0243	0,0235	0,0234	0,0232	**0,0230**
SEN	0,0338	0,0345	0,0342	0,0337	**0,0310**
ALE	0,0271	0,0265	0,0265	0,0263	**0,0260**
AGR	0,0176	0,0169	0,0169	0,0168	**0,0158**
TRI	0,0390	0,0389	0,0388	0,0385	**0,0381**

Tabla 5.18: RMSE relativo de la energía por estilo con CBR

Estilo	RMSE relativo
NEU	0.64
SEN	0.91
ALE	0.68
AGR	0.69
TRI	0.68

5.4. Evaluación subjetiva

Las medidas objetivas de evaluación de un sistema basado en corpus conllevan de forma implícita la comparación con locuciones concretas de un hablante en un momento determinado. Sin embargo, existe más de un modo de pronunciar aceptablemente una frase y, además, los oyentes pueden tener diferentes preferencias. Por lo tanto, se hace necesaria la realización de una evaluación subjetiva mediante una prueba de percepción en la cual los oyentes manifiesten sus preferencias o puntúen los estímulos presentados (Llisterri et al., 1999).

En el ámbito de la síntesis del habla se pueden evaluar diferentes características como son la inteligibilidad, la naturalidad y la expresividad. En algunas aplicaciones, como por ejemplo, en las máquinas parlantes para personas invidentes, la inteligibilidad del habla a alta velocidad es más importante que la naturalidad (Llisterri et al., 1993). En cambio, una prosodia correcta y una elevada naturalidad son esenciales en la mayoría de aplicaciones multimedia (Lemmetty, 1999). La evaluación se puede realizar a diferentes niveles (segmento, palabra, frase o párrafo) y con diferentes tipos de pruebas (Campbell, 2007).

En el presente trabajo sobre estimación de la prosodia en el ámbito de la síntesis del habla expresiva, se requiere una evaluación en el nivel de frase, que permita valorar la capacidad de aprendizaje del sistema automático para generar la prosodia de un texto oralizado en un estilo expresivo concreto.

5.4.1. Preparación de los estímulos

Para evaluar la predicción de la prosodia se ha entrenado el sistema con el 75 % de las frases del corpus; del 25 % restante se han escogido 15 frases de cada estilo, que son las empleadas en la evaluación. Para cada frase de test se generan dos estímulos: uno con la prosodia sintética (PS) estimada a partir del texto, y otro con la prosodia natural (PN) extraída a partir del etiquetado del corpus. Los ficheros de prosodia que constituirán la entrada del sintetizador de voz contienen la transcripción fonética y los valores de duración en ms, de energía *rms* y de F_0 en Hz para cada segmento. Un ejemplo de este tipo de entrada se puede observar en la figura 5.21.

A partir de la información prosódica —natural o sintética—, se generan los archivos de audio correspondientes mediante el sistema de síntesis del habla basado en selección de unidades desarrollado por el GPMM de EALS-URL (Alías et al., 2005). Dicho sistema de síntesis esta configurado para minimizar el número de puntos de concatenación, lo que tiende a priorizar la selección de unidades correlativas. Por lo tanto, únicamente se establece un coste de concatenación binario, que vale 1 si los dos difonemas (unidad mínima) que se concatenan son consecutivos en la misma frase del corpus, o 0 en caso contrario. Con el objetivo de centrar únicamente la evaluación en la prosodia y no en el procesamiento de la señal, las frases de test forman parte de la base de datos del sintetizador, aunque, como se ha señalado, no se utilizaron en el entrenamiento del módulo prosódico. De esta forma,

```
_       500     0.0002  148
p       50      0.0083  146
O       105     0.0500  149
r       45      0.0179  155
m       95      0.1045  161
A       220     0.1021  186
R       125     0.0079  205
_       595     0.0004  163
E       90      0.0555  134
l       80      0.0623  137
B       35      0.0053  141
j       140     0.0764  145
A       135     0.0544  148
x       95      0.0055  183
e       40      0.0659  202
E       85      0.0513  181
s       105     0.0065  148
O       95      0.0250  144
t       65      0.0157  140
r       50      0.0652  140
a       75      0.0811  135
k       105     0.0083  135
O       100     0.0753  139
s       120     0.0147  133
a       120     0.0347  140
_       500     0.0006  139
```

Figura 5.21: Ejemplo de fichero de prosodia de la frase *Por mar, el viaje es otra cosa*. La primera columna corresponde a la transcripción fonética, la segunda a la duración en ms, la tercera a la energía *rms* y la cuarta a la F_0 en Hz.

tanto para los estímulos PN como para los estímulos PS se parte de frases enteras que se procesarán para ajustar la prosodia a los valores de entrada. Este proceso de re-síntesis se realiza en el nivel del difonema o del trifonema ajustando los valores de duración y de F_0 mediante una técnica basada en TD-PSOLA (Moulines y Charpentier, 1990) descrita en Iriondo et al. (2003). La energía se ajusta en el nivel segmental mediante una función de ganancia aplicada directamente sobre las muestras de la señal de voz.

Del entrenamiento prosódico realizado para cada estilo mediante una validación cruzada en 4 bloques, se ha escogido uno de los cuatro bloques, de forma que se ha dispuesto de un 75 % de las frases para el entrenamiento y del 25 % restante para el test. De cada subconjunto de test (uno por estilo) se han escogido 15 frases para la prueba subjetiva. La elección de estas frases se ha basado en dos criterios:

- Utilizar frases de longitud cercana a la media (en número de GA y de segmentos) de cada corpus.
- Disponer de frases con un amplio rango de valores del RMSE de F_0 para posibilitar la comparación de los resultados de la pruebas objetiva y subjetiva, aunque solo sea desde el punto de vista de la melodía. Por el momento, no se ha establecido una medida global objetiva que incluya la evaluación de los tres parámetros prosódicos simultáneamente, por lo que se ha escogido el que a priori parece más relevante en el ámbito de la expresividad oral.

Con este fin se han calculado los tres cuartiles[11] según el valor del RMSE para la

[11] El primer cuartil (Q1) se define como la mediana de la primera mitad de valores; el segundo cuartil

F_0 para cada estilo y se han seleccionado unas 20 frases alrededor de cada cuartil. De esta preselección se han escogido definitivamente las 5 frases que más se aproximan a la media de la duración (en número de GA y de segmentos) del estilo correspondiente. La tabla 5.19 muestra los valores de los cuartiles del RMSE de la F_0 y los promedios de duración de las frases de cadas estilo. Esta distribución de las frases permitirá estudiar si existe relación entre la percepción subjetiva y el error cometido en la predicción de la F_0. En la tabla 5.20 se muestra el promedio de los valores de RMSE y de ρ para la F_0, la duración y la energía de las frases seleccionadas para el test subjetivo.

Tabla 5.19: Cuartiles del RMSE para la F_0, junto con el promedio del número de GA y segmentos, del subconjunto de frases de test que ha servido de base para la preselección y la selección definitiva de las frases de la prueba subjetiva.

	Q1	Q2	Q3	Núm. de GA	Núm. de segmentos
NEU	23,3	30,3	37,8	4,4	31
SEN	18,6	22,5	26,8	3,6	24
ALE	62,3	72,5	83,5	3,8	27
AGR	42,5	50,1	58,1	5,9	38
TRI	16,5	21,8	27,3	4,6	32

Tabla 5.20: Valores promedio de RMSE y de ρ en los tres parámetros prosódicos de las frases que forman la prueba subjetiva.

	F0 (Hz)		Duración (ms)		Energía (rms)	
Estilo	RMSE	ρ	RMSE	ρ	RMSE	ρ
NEU	30,76	0,60	19,92	0,76	0,024	0,81
SEN	22,68	0,50	26,59	0,70	0,027	0,72
ALE	72,86	0,56	24,05	0,78	0,024	0,76
AGR	50,45	0,69	30,54	0,66	0,017	0,76
TRI	22,18	0,57	36,06	0,57	0,038	0,77

Las figuras 5.22, 5.23, 5.24, 5.25 y 5.26 muestran un ejemplo para cada estilo de los valores de la PS obtenidos para una frase de test y de los valores de la PN en la misma frase del corpus.

(Q2) como la propia mediana de la serie; el tercer cuartil (Q3) se corresponde a la mediana de la segunda mitad de valores.

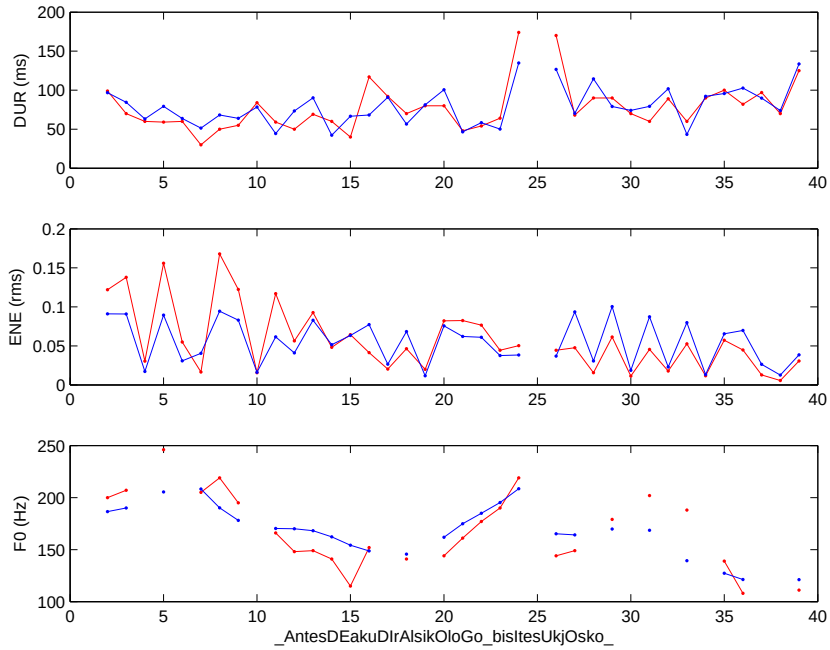


Figura 5.22: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase *Antes de acudir al psicólogo, visite su quiosco* en estilo neutro.

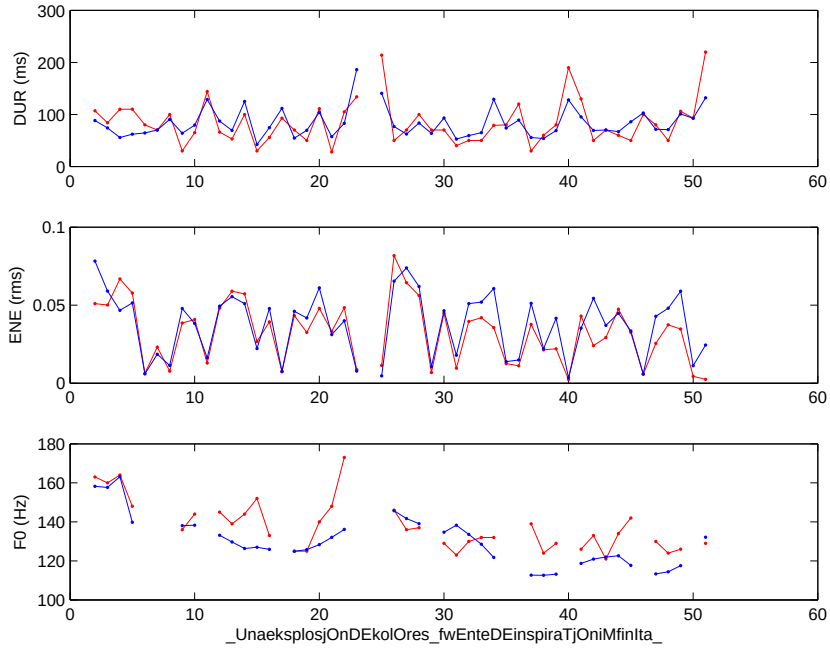


Figura 5.23: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase *Una explosión de colores, fuente de inspiración infinita* en estilo sensual.

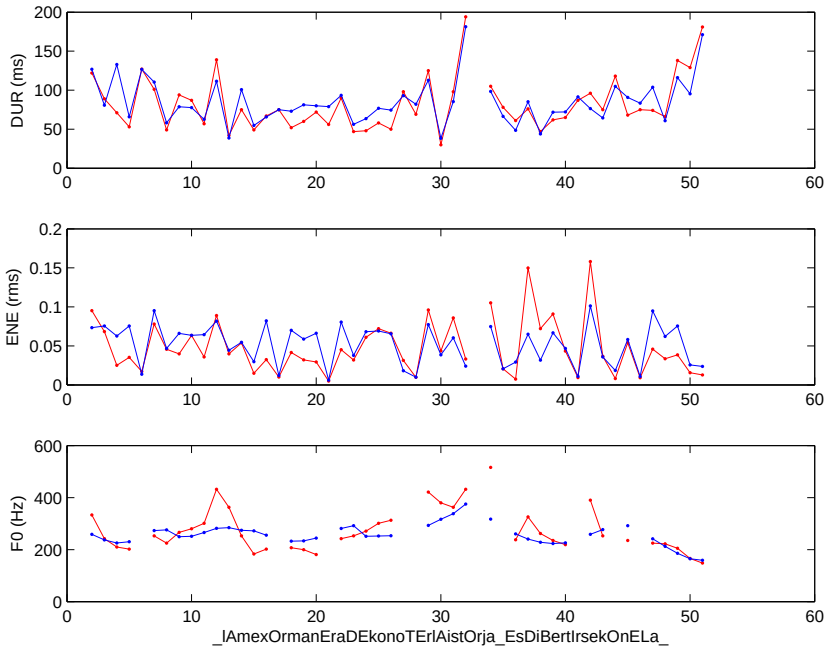


Figura 5.24: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase *Trescientos millones, cambian la vida.* en estilo alegre.

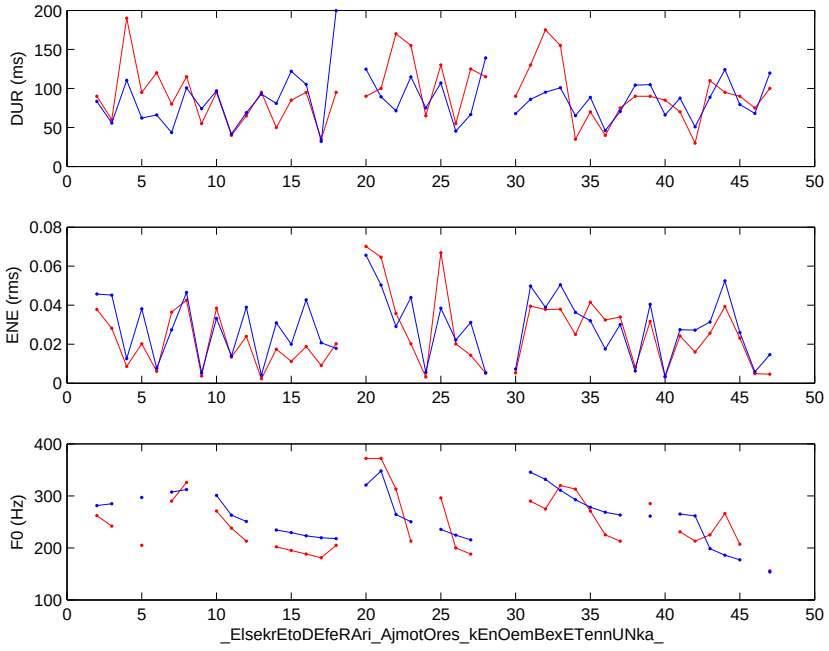


Figura 5.25: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase *El secreto de Ferrari. Hay motores, que no envejecen nunca.* en estilo agresivo.

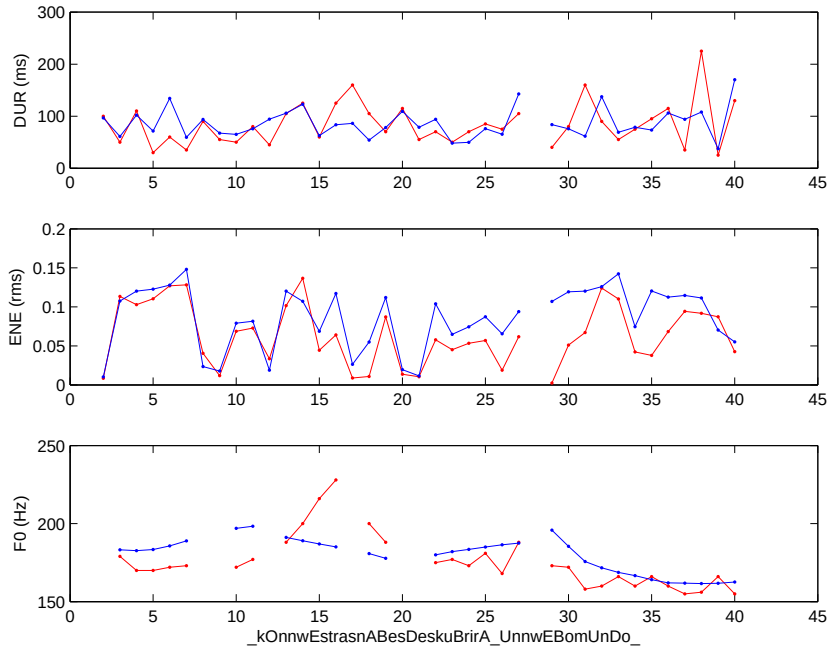


Figura 5.26: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase *Con nuestras naves descubrirá, un nuevo mundo.* en estilo triste.

En el anexo D se muestra, para cada frase de la prueba subjetiva, la siguiente información: el texto, tres gráficas con los valores de PN y PS de F_0, duración y energía y, por último, las medidas objetivas de RMSE y de ρ.

5.4.2. Pruebas perceptivas

Una vez se han preparado los estímulos, se debe decidir el tipo de prueba más adecuado para presentarlos a los oyentes y la metodología de evaluación de los mismos. El objetivo de la prueba es, como ya se ha señalado, la evaluación de la generación automática de la prosodia para cada estilo expresivo. Se dispone de una pareja de ficheros de sonido por cada frase que se debe evaluar, fruto de la resíntesis con PN (copiada del etiquetado del corpus) y con PS (predicha a partir del texto). Por ello, se plantean diferentes posibilidades de presentación de los estímulos (de forma individual o por parejas) y de escalas de puntuación. A partir de la recomendación P.800 de la Unión Internacional de Telecomunicaciones (UIT) (UIT-T, 1996), se consideran tres posibles métodos de prueba perceptiva que podrían ser adecuadas para este caso:

1. Determinación de índices por categorías absolutas —*Absolute Category Rating*— (ACR) obteniéndose una nota media de opinión —*Mean Opinion Score*— (MOS).

2. Determinación de índices por categorías de degradación —*Degradation Category Rating*— (DCR) obteniéndose una nota media de opinión sobre las degradaciones —

Degradation Mean Opinion Score— (DMOS).

3. Determinación de índices por categorías de comparación —*Comparison Category Rating*— (CCR) obteniéndose una nota media de opinión sobre las comparaciones —*Comparison Mean Opinion Score*— (CMOS).

A continuación se describen brevemente los aspectos principales de estos métodos de determinación subjetiva de la calidad de transmisión y su adaptación a la evaluación de la prosodia.

5.4.2.1. Método de determinación de índices por categorías absolutas

Se trata del método de prueba de escucha que se basa en "índices de categorías absolutas" (ACR, *absolute category rating*) y utiliza una escala de evaluación de 5 notas. En la recomendación se recalca la importancia de la disposición y del enunciado de las escalas de opinión, por lo que estas deben seguir las normas a las que se ha llegado tras muchos años de experiencia. La escala más utilizada es:

Excelente	5
Buena	4
Regular	3
Mediocre	2
Mala	1

La medida MOS representa la magnitud promediada a partir de las notas de varios usuarios al evaluar diferentes estímulos.

La utilización de este método para la evaluación del módulo prosódico desarrollado se puede llevar a cabo mediante la presentación de los 30 estímulos de cada estilo de forma individual y en orden aleatorio. El oyente no realiza una comparación directa para cada pareja de estímulos PN y PS, sino que esta comparación quedará implícita en los resultados y se podrá obtener mediante un análisis posterior de los mismos. El mayor inconveniente de este método puede ser la dificultad de mantener un criterio común sobre el significado de las notas al aplicarlas a la evaluación de la prosodia.

5.4.2.2. Método de determinación de índices por categorías de degradación

Se trata de una variación del método ACR, indicada para comparar parejas de estímulos de elevada calidad y en la que las diferencias pueden ser difíciles de detectar a posteriori. El orden de las parejas se mantiene constante, presentándose en primer lugar la muestra de referencia de calidad alta y, en segundo lugar, la muestra correspondiente sometida a evaluación.

La escala propuesta en la recomendación es la siguiente:

Degradación inaudible	5
Degradación audible, pero no molesta	4
Degradación ligeramente molesta	3
Degradación molesta	2
Degradación muy molesta	1

La magnitud obtenida a partir del promedio de las notas (nota media de opinión sobre las degradaciones) se representa por la medida DMOS.

La utilización de esta prueba para la evaluación de la prosodia a partir de la comparación de las mismas frases con PN y con PS puede ser adecuada si se sustituye el concepto de degradación por el de similitud o parecido en la calidad. Por lo tanto, se pueden modificar las categorías de degradación por unas categorías de similitud (p. ej.: *Similitud muy alta*, *Similitud alta*, *Cierta similitud*, *Similitud baja* y *Ninguna similitud*). Esta adaptación de la recomendación DCR se asemeja al método CCR que se describe a continuación.

5.4.2.3. Método de determinación de índices por categorías de comparación

Este método es parecido al DCR pero, a diferencia de éste, en el procedimiento CCR se elige al azar en cada prueba el orden de las muestras procesada y no procesada.

Los oyentes utilizan la escala siguiente para calificar la calidad de la segunda muestra con relación a la de la primera:

Mucho mejor	3
Mejor	2
Ligeramente mejor	1
Aproximadamente igual	0
Ligeramente peor	-1
Peor	-2
Mucho peor	-3

La cantidad obtenida a partir del promedio de las puntuaciones (nota media de opinión sobre las comparaciones) viene representada por la medida CMOS.

Una posible ventaja del método CCR sobre el DCR es la posibilidad de evaluar el procesamiento de la señal vocal, que o bien degrada o bien mejora la calidad de la voz.

Para el caso de la evaluación de la prosodia, la muestra procesada equivaldría a la versión con PS y la muestra no procesada al estímulo con PN.

5.4.3. Elección del tipo de prueba

Los tres tipos de pruebas presentados en el apartado anterior pueden ser válidos para la evaluación que se va a llevar a cabo, pero la realización de las tres para cada estilo

supondría un total de 15 pruebas para cada evaluador. Según UIT-T (1996), idealmente ninguna sesión debe durar más de 20 minutos y en ningún caso debe rebasar los 45 minutos. La duración de una prueba con 30 frases individuales o con 15 parejas puede oscilar entre 4 y 6 minutos. Por lo tanto, es aconsejable realizar únicamente una prueba por estilo de forma que completar la prueba entera costaría entre 20 y 30 minutos, sin contar el tiempo de descanso entre estilos. Duplicar o triplicar el número de pruebas con la finalidad de tener más resultados podría resultar contraproducente, ya que pocos usuarios terminarían todo el experimento. Una vez decidido que únicamente se realizaría un tipo de prueba, faltaba determinar cuál.

Con esta finalidad, se preparó una interfaz de test para cada tipo de prueba y se propuso a un grupo de siete expertos en el ámbito de las tecnologías del habla, concretamente del GPMM y del *Grup de Fonètica del Departament de Filologia Espanyola de la Universitat Autònoma de Barcelona*, que realizaran los tres tests para decidir el método concreto que se emplearía con un grupo mayor de oyentes a partir de sus comentarios y de los resultados de esta prueba piloto. Finalmente, el método escogido fue el ACR, que proporciona una medida MOS (véase el apartado 5.4.2.1). Las principales razones de esta elección se describen a continuación:

1. Los enunciados con PN, además de tener en general una mejor pronunciación, también presentan una mejor calidad segmental que los enunciados con PS. Esto es debido al proceso de resíntesis basado en TD-PSOLA, ya que en los primeros prácticamente no hay modificación de la señal. En cambio, para los segundos, en general, las modificaciones de F_0 y de duración segmental son mayores. Por lo tanto, en los enunciados con PS pueden aparecer errores de carácter segmental que reducen su calidad global. Aunque la pregunta se centre únicamente en la evaluación de la prosodia, es difícil abstraerse de esta pérdida de calidad y centrarse únicamente en aspectos como la entonación, el ritmo o el énfasis. Si los enunciados se muestran por parejas (DCR o CCR), esta diferencia se acentúa más que si se escuchan de forma independiente en momentos distintos (ACR).

2. La comparación directa de dos formas prosódicas para un mismo enunciado presenta cierta complejidad, ya que dos realizaciones diferentes pueden ser adecuadas para un texto y un estilo determinados. La PN de una locutora profesional tiene que estar prácticamente siempre cerca de la calidad máxima. La evaluación de la PS generada de forma automática se tiene que realizar en términos de similitud en la calidad respecto de la PN. Si los dos estímulos se presentan simultáneamente, la PN puede condicionar la respuesta del oyente en el sentido que la perciba como la única manera correcta de decir el texto correspondiente. Un enunciado que inste al oyente a valorar el parecido de los dos estímulos en términos de la calidad de la prosodia sería demasiado complicado para evaluadores no expertos.

3. Por último, el análisis de los resultados permite llegar a un mayor nivel de profundidad con el método ACR, ya que se dispondrá de una nota MOS para el grupo de estímulos con PN y de otra para el grupo de estímulos con PS. Sin embargo, los métodos DCR o CCR únicamente proporcionan una nota relativa que solo permitiría comparaciones relativas entre estilos. Además, con el método ACR se puede obtener

una medida comparativa restando las notas de cada pareja de estímulos para cada usuario y, aplicando un escalado adecuado, disponer de una medida del tipo CMOS o DMOS.

A partir de los comentarios de los expertos y siguiendo las recomendaciones presentes en UIT-T (1996), se ha redactado una página inicial con una explicación más detallada de los aspectos que los oyentes deben valorar utilizando un lenguaje más llano y sin tecnicismos. Además, para cada estilo se han incorporado en esta página inicial seis estímulos que no forman parte de la evaluación para que el oyente se familiarice con la voz sintética. Por último, también se ha simplificado la pregunta que se presenta al evaluador junto con el estímulo que debe puntuar. Estos textos se pueden consultar en el anexo D.

Otro aspecto importante es la traducción de la escala MOS del inglés (*Excellent (5), Good (4), Fair (3), Poor (2), Bad (1)*) al castellano. Las recomendaciones de la UIT P.85 (UIT-T, 1994) y P.800 (UIT-T, 1996) difieren en la traducción de *Fair*, ya que sugieren *Pasable* y *Regular* respectivamente. La primera traducción nos ha parecido un término un poco coloquial y, finalmente, hemos optado por el segundo que, además, corresponde al que se encuentra en la recomendación más reciente.

5.4.4. Realización de la prueba y resultados

Se han preparado cinco pruebas diferentes, una para cada estilo, accesibles en la web mediante la autentificación con una dirección de correo electrónico. Cada prueba se puede suspender en cualquier momento y reemprenderla posteriormente. Los evaluadores son alumnos y profesores de la universidad a los que se solicitó su colaboración por correo electrónico.

Únicamente se han tenido en consideración aquellos evaluadores que han completado las cinco pruebas para poder comparar los resultados entre estilos, concretamente 18 hombres y 12 mujeres con edades comprendidas entre los 20 y los 44 años. La lengua materna de todos ellos es el castellano o el catalán.

Para el análisis estadístico y la presentación de los resultados también se han seguido las indicaciones dadas en las recomendaciones UIT-T (1994) y UIT-T (1996).

En primer lugar se ha calculado los valores MOS obtenidos para ambos conjuntos de estímulos (PN y PS), distinguiendo las notas de los participantes masculinos y femeninos, ya que la voz evaluada se corresponde a la de una mujer. En la tabla 5.21 se muestran los resultados para cada estilo y la representación gráfica de los valores globales se puede observar en la figura 5.27. No se aprecia ninguna tendencia que indique un comportamiento distinto en función del sexo de los evaluadores.

El MOS de la PN servirá de referencia para poder evaluar el funcionamiento del sistema de generación automática de prosodia. En primer lugar, se constata (figura 5.27) que se obtienen valores diferentes para cada estilo: el valor máximo se alcanza en el estilo ALE (4.35), NEU y AGR presentan valores intermedios (4,01 y 3,96) y, finalmente, los

Tabla 5.21: Valores MOS para los estímulos con PN y con PS para cada estilo, distinguiéndose los resultados de los participantes masculinos (H) y de los femeninos (M).

	NEU	SEN	ALE	AGR	TRI
PN	**4,01**	**3,69**	**4,35**	**3,96**	**3,70**
PN (H)	4,07	3,74	4,38	4,03	3,81
PN (M)	3,93	3,60	4,30	3,85	3,53
PS	**3,14**	**3,14**	**3,12**	**2,70**	**3,44**
PS (H)	3,14	3,23	3,16	2,64	3,46
PS (M)	3,14	3,01	3,05	2,78	3,40

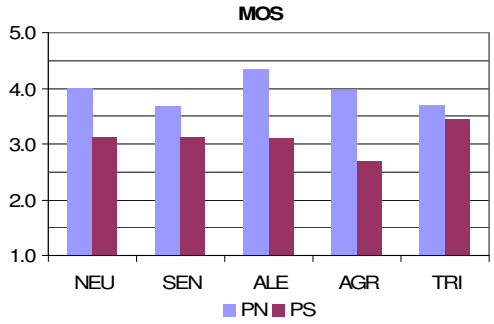


Figura 5.27: Valores MOS para los estímulos con PN y con PS para cada estilo

estilos TRI y SEN obtienen la peor puntuación (3,70 y 3,69). También se ha realizado un análisis de varianza —*ANalysis Of VAriance*— (ANOVA) con las 5 muestras de notas con PN y una prueba de comparación múltiple de diferencias enteramente significativas (HSD, *Honestly Significant Difference*) o prueba de Tukey (Tukey, 1953, citado por UIT-T, 1996, p. 25). El resultado de esta prueba indica que la diferencia de medias entre estos tres grupos es significativa (véase la gráfica 5.28e). Es decir, el MOS de ALE es significativamente diferente de los otros cuatro, los valores MOS de NEU y AGR son significativamente diferentes de los otros tres pero no entre ellos y lo mismo ocurre en los estilos TRI y SEN. Esta diferencia de medias entre algunos estilos, aunque todas las frases tengan una PN, puede deberse a que el proceso de resíntesis genera una calidad final diferente en función de las características de la voz propias de cada estilo. El análisis de los resultados sobre las preferencias de nota muestran que más del 60 % se corresponden a *Excelente* y *Buena* para todos los estilos, alcanzando el 86,9 % para el estilo ALE (véase el histograma apilado de la figura 5.28a). En cambio, la suma de puntuaciones *Mala* y *Mediocre* se sitúa por debajo del 15 % para todos los estilos.

Los resultados de MOS globales obtenidos para la PS (figura 5.27) muestran que el estilo TRI obtiene mayor puntuación (3,44). El MOS de AGR es el menor de todos (2,70) y los estilos NEU, SEN y ALE consiguen prácticamente la misma puntuación ($\sim$ 3,14). Sobre estos valores medios, el ANOVA y la comparación múltiple (véase la gráfica 5.28f) permiten afirmar que tanto el MOS de TRI como el de AGR son significativamente diferentes del resto y viceversa. El análisis mediante un histograma acumulado de las puntuaciones por cada estilo (figura 5.28b) muestra que sobre el 40 % de las puntuaciones se corresponden a *Excelente* y *Buena* para NEU, SEN y ALE, superándose el 50 % para TRI, pero situándose

justo por debajo del 20 % para AGR. Si se incluye en estos resultados el porcentaje de respuestas *Regular*, se alcanza prácticamente el 75 % para todos los estilos menos para AGR (60,95 %). La respuesta *Mala* tiene diferentes comportamientos en función del estilo: inferior al 5 % para NEU y TRI, inferior al 9 % para SEN y ALE, y de casi el 13 % para AGR.

También se han representado las distribuciones acumuladas para la PN (figura 5.28c) y la PS (figura 5.28d). En este tipo de representación las curvas que van por debajo corresponden a mejores puntuaciones. En el caso de la PN, se observa claramente el dominio del estilo ALE y comportamientos parecidos en las parejas AGR - NEU y SEN - TRI. Para la PS, se confirma la agrupación de los estilos NEU, SEN y ALE en una zona intermedia entre el estilo con una mejor puntuación (TRI) y el peor valorado (AGR).

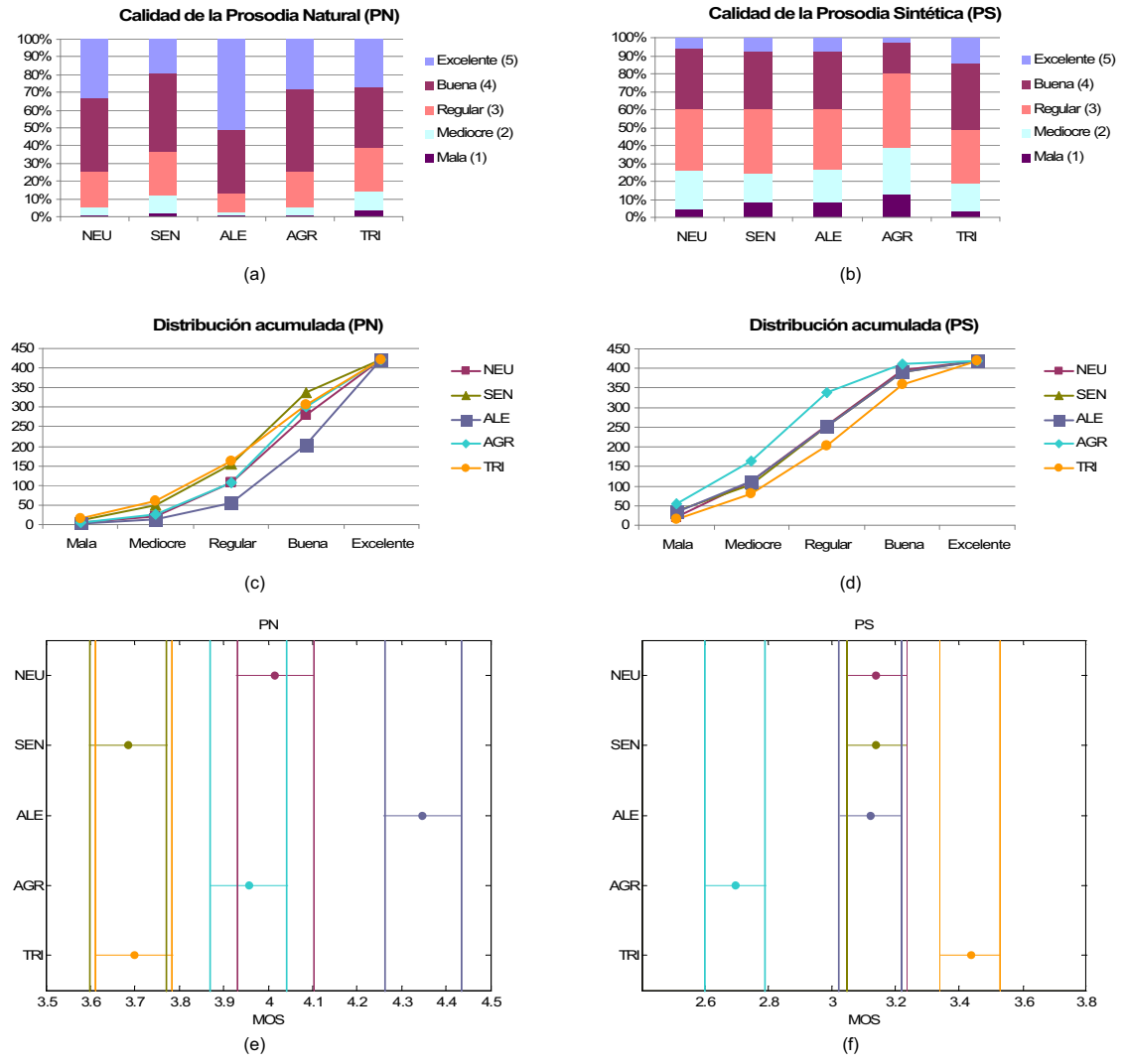


Figura 5.28: Comparación de los resultados de la prueba ACR para los estímulos con PN y con PS: (a) y (b) Histogramas apilados en porcentaje; (c) y (d) Distribuciones acumuladas; (e) y (f) Resultado de las comparaciones múltiples HSD.

Una vez analizados los resultados de las frases con PN y con PS por separado, conviene comparar, para cada estilo, los resultados obtenidos con la PS con los de las frases con PN. Un primer análisis se puede llevar a cabo mediante la diferencia de valores MOS. Se han realizado pruebas de significación mediante técnicas de ANOVA para cada pareja de valores MOS PN-PS obtenidas por estilo, rechazándose para todos los casos la hipótesis nula de que ambas medias son iguales. La figura 5.29 muestra un diagrama de cajas (en inglés *boxplot*) realizado a partir de todas las notas de las dos versiones PN y PS de cada estilo. Cada pareja de cajas nos da una idea de cada distribución de notas según el estilo y el tipo de prosodia. Además, se han añadido los 10 valores MOS, representados con el símbolo μ. La diferencia de medias da el siguiente orden de mejor (menor diferencia) a peor (mayor diferencia): TRI (0,26), SEN (0,55), NEU (0,88), ALE (1,23) y AGR (1,26).

Los diagramas de cajas de la pareja TRI y la pareja SEN también muestran un mayor parecido que el del resto de estilos. Aunque las cajas del NEU-PS y el ALE-PS son iguales, la diferencia de medias con sus respectivas versiones PN son diferentes debido a que el MOS de ALE-PN es mayor que el de NEU-PN. La PS del estilo AGR presenta el peor resultado tanto relativo (mayor diferencia de medias) como absoluto (menor valor de MOS).

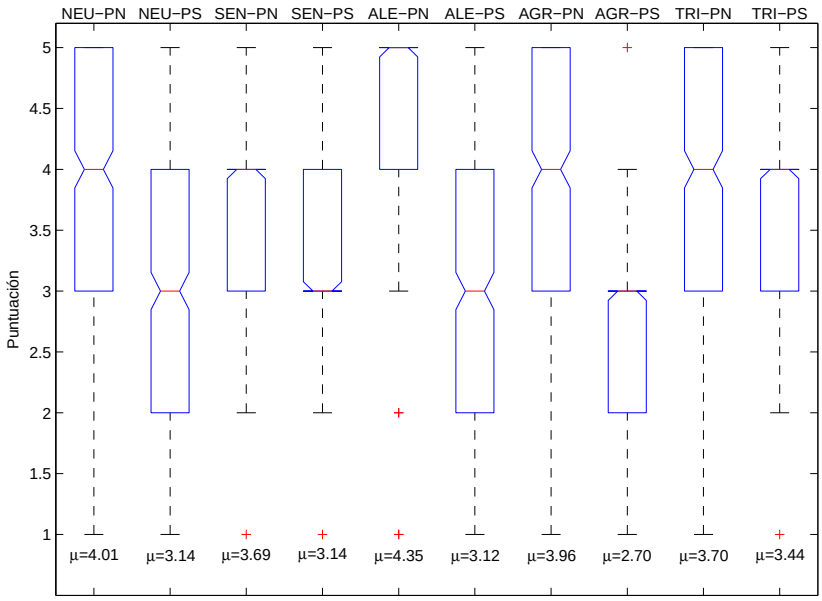


Figura 5.29: Diagrama de cajas realizado a partir de las puntuaciones de cada estilo con PN y con PS. Se incluye el valor MOS de cada categoría, representado por μ.

Un análisis comparativo más detallado puede realizarse si se calcula el valor de la diferencia entre las puntuaciones individuales de cada par PN-PS correspondiente a una frase y a un oyente. Con el objetivo de tener una medida DMOS para medir la similitud de dos tipos de prosodia en términos de calidad (adecuación al estilo pretendido), tal y

como se ha propuesto en el apartado 5.4.2.2, se pueden aplicar las fórmulas siguientes:

$$dif = n_{PN} - n_{PS} \tag{5.11}$$

$$sim = 5 - max(dif, 0) \tag{5.12}$$

donde n_{PN} y n_{PS} son las puntuaciones asignadas por un participante en el test para las dos versiones PN y PS de una misma frase respectivamente. Por tanto, para cada pareja de locuciones PN-PS evaluadas se dispone de una medida de similitud (*sim*) escalada entre 1 (Ninguna similitud) y 5 (Similitud muy alta). En aquellos casos en los que la puntuación de la PS supera a la de su pareja PN, no se permite que el resultado sea mayor que 5, limitándose a este valor. Se han contabilizado los siguientes porcentajes de casos en los que es necesario aplicar esta limitación: 13 % para NEU, 16 % para SEN, 4 % para ALE y AGR, y 23 % para TRI. Prescindiendo de esta restricción, el valor de DMOS final aumentaría para todos los estilos, pero se considera más correcto que el DMOS represente valores que pertenecen al mismo rango de 1 a 5.

Finalmente, se ha definido la medida DMOS como el promedio de valores *sim* (ecuación 5.12) pertenecientes a un mismo grupo (p. ej. las notas de todas las frase de un estilo). La figura 5.30 representa el DMOS obtenido para cada estilo. El orden de mejor (mayor DMOS) a peor (menor DMOS) se mantiene igual que el obtenido anteriormente mediante la diferencia de valores MOS. TRI y SEN, con 4,44 y 4,26 respectivamente, son los estilos que presentan mayor parecido, seguidos por NEU con 3,97 y, finalmente, ALE y AGR con 3,73 y 3,69 respectivamente.

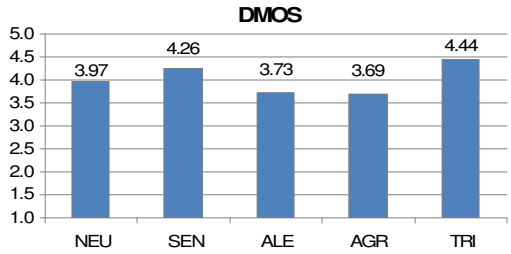


Figura 5.30: DMOS obtenido a partir de las puntuaciones individuales de cada par de frases con PN y con PS.

Las gráficas mostradas en la figura 5.31 permiten un análisis más detallado de estos resultados. Los niveles de similitud *Muy Alta* y *Alta* superan el 60 % para todos los estilos y, en concreto, son mayores del 80 % para TRI y SEN (véase el histograma apilado de la figura 5.31a). La distribución acumulada (figura 5.31b) muestra claramente el comportamiento de los 5 estilos, donde destaca la curva con mayor concavidad del estilo TRI. La figura 5.31c es el resultado de una comparación múltiple HSD basada en un ANOVA que muestra que el DMOS del NEU es significativamente diferente de los cuatro DMOS restantes. Los valores DMOS de ALE y AGR se superponen claramente y, muy ligeramente, los de SEN y TRI. Un diagrama de cajas (figura 5.32) obtenido a partir de estos datos, permite observar que las medianas de todos los estilos presentan un valor de 4 (*Alta*), a excepción de la mediana de TRI, que alcanza un valor de 5 (*Muy alta*).

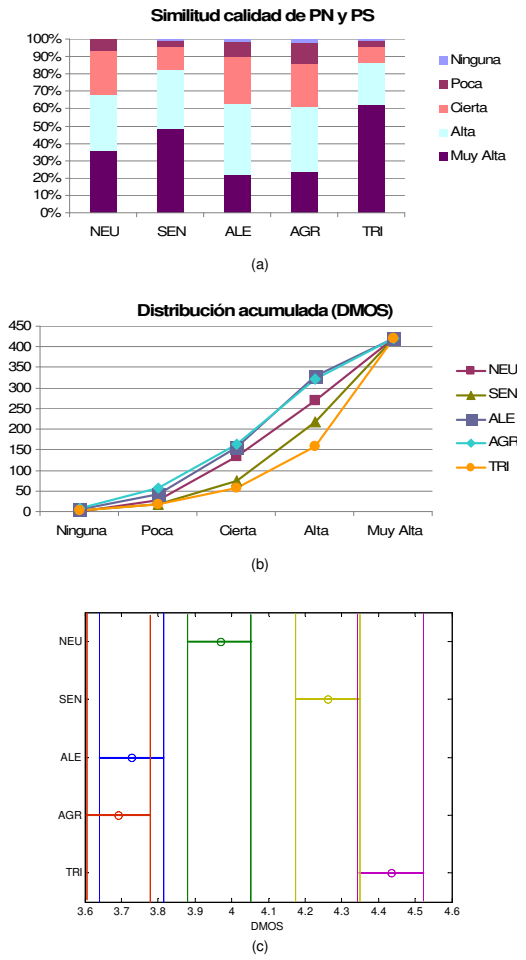


Figura 5.31: Resultado del análisis comparativo de la PN y la PS: (a) Histograma apilado en porcentaje; (b) Distribución acumulada; (c) Resultado de la comparación múltiple.

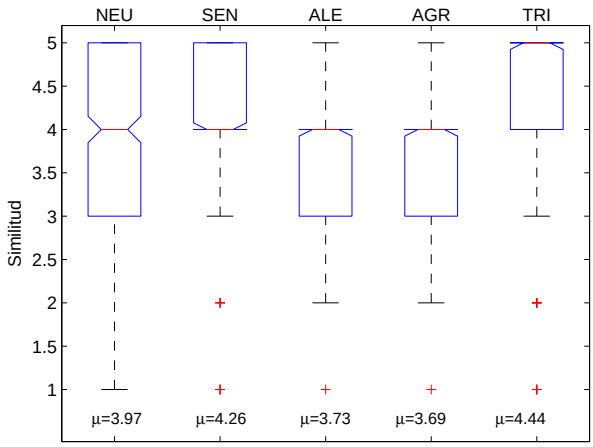


Figura 5.32: Diagrama de cajas a partir de las puntuaciones de similitud entre la PN y la PS de los estímulos de cada estilo. Se incluye también el valor DMOS de cada estilo, representado por μ.

Por último se han analizado los resultados del DMOS de los estímulos agrupados según la cercanía de sus valores RMSE de la predicción de F_0 a los cuartiles de esta medida para todo el conjunto de test (véase el apartado 5.4.1). Los resultados para cada estilo y el total para los cinco estilos se muestran en la tabla 5.22. En la columna Q1 se muestra el resultado DMOS para los estímulos próximos al primer cuartil, es decir el de menor valor de RMSE. Q2 representa la mediana y Q3 el tercer cuartil. A la vista de los resultados, no existe ninguna evidencia de que los oyentes muestren cierta preferencia por aquellos estímulos que han obtenido menor RMSE en la estimación de la F_0.

Tabla 5.22: Valores DMOS obtenido a partir de la comparación de los estímulos con PN y con PS para cada estilo y total, distinguiéndose los resultados en función de la proximidad a los tres cuartiles del RMSE de la F_0.

Estilo	Q1	Q2	Q3
NEU	3,96	3,92	4,02
SEN	4,23	4,16	4,43
ALE	3,74	3,81	3,64
AGR	3,83	3,67	3,60
TRI	4,39	4,43	4,49
Total	4,03	4,00	4,04

El análisis de los resultados de esta prueba subjetiva ha permitido descubrir un orden de preferencia según el estilo: triste, sensual, neutro, alegre y agresivo (de mayor a menor). Otros experimentos del presente trabajo de tesis han revelado prácticamente el mismo orden en la identificación subjetiva de estilos a partir de muestras grabadas (véase el apartado 4.3). Si nos fijamos en las medidas utilizadas para evaluar objetivamente la proximidad de la PS con respecto a la PN, comprobaremos que también se obtiene el mismo orden para el RMSE de la F_0 de las frases utilizadas en la prueba subjetiva (véase la tabla 5.20). La única diferencia entre ambos casos es el orden de AGR y ALE, pero hay que tener en cuenta que la diferencia en las medidas DMOS de ambos no es significativa. Esta comparación, muestra que en la percepción subjetiva del habla expresiva influyen aspectos relacionados con el modelado prosódico, sobre todo relacionados con la melodía. Sin embargo, las diferencias de valores MOS de las frases con PN surgidas al comparar los diferentes estilos indican que su definición o la simulación realizada por parte de la locutora también pueden influir en los oyentes.

5.5. Resumen

En este capítulo se han detallado las diferentes fases del desarrollo y evaluación de un módulo de generación automática de la prosodia en el ámbito de la síntesis del habla expresiva utilizando el corpus oral descrito en el capítulo 4. El principio del capítulo presenta dos investigaciones preliminares que han servido de base para el desarrollo del sistema final (Iriondo et al., 2000, 2004).

De acuerdo con el estado de la cuestión, la experiencia adquirida y los retos que se deseaban abordar, se ha desarrollado un sistema de estimación de la prosodia basado en corpus que se caracteriza por modelar de forma conjunta las funciones lingüística y paralingüística de la prosodia, a partir de la extracción automática de atributos prosódicos del texto, que son la entrada de un sistema de aprendizaje automático que predice los rasgos prosódicos modelados previamente (Iriondo et al., 2006, 2007d).

El sistema de modelado prosódico presentado en este trabajo se fundamenta en el razonamiento basado en casos —una técnica de aprendizaje automático por analogía. Dicho sistema se ha desarrollado por completo en el ámbito de la presente tesis. La comparación con otros métodos de aprendizaje automático ya implementados en Weka ha sido satisfactoria. Las ventajas de desarrollar un método propio, en lugar de utilizar *software* ajeno, es el control total de todos los algoritmos y variables, la libre utilización en aplicaciones para terceros y la posibilidad de sofisticar módulos internos.

Para el ajuste de algunos parámetros del sistema desarrollado y se evaluación se han utilizado medidas objetivas del error y la correlación calculados en las locuciones de del conjunto de test. En todas las pruebas se ha realizado una validación cruzada en bloques que garantiza que todas las muestras del corpus forman parte una vez del conjunto de test.

Dado que las medidas objetivas siempre son respecto casos concretos, no aportan información sobre el grado de aceptación que tendrá el habla sintetizada en los oyentes. Por lo tanto, se han llevado a cabo una serie de pruebas de percepción cuyos participantes han puntuado una serie de estímulos de cada estilo. Se ha realizado un estudio y unas pruebas con expertos para definir el tipo de prueba considerada más idónea para proponer al público en general.

Finalmente, los resultados se han analizado para cada estilo y se han comparado con las medidas objetivas obtenidas, lo que ha permitido extraer algunas conclusiones sobre la relevancia de los rasgos prosódicos en el habla expresiva y constatar que los resultados generados por el módulo prosódico han tenido una buena aceptación, aunque no por igual para todos los estilos.

Capítulo 6

Conclusiones y futuras líneas de investigación

Las conclusiones del trabajo y las propuestas de futuras líneas de investigación se presentan organizadas en cuatro apartados. Un primer apartado que constituye una exposición general del trabajo (apartado 6.1), otros dos apartados que describen con mayor detalle aspectos concretos de las dos contribuciones principales de la presente tesis: el corpus oral (apartado 6.2) y el modelado de la prosodia (apartado 6.3); y un último apartado en el que, a partir de la experiencia y el conocimiento adquirido, se proponen diferentes líneas de investigación en el campo de la síntesis del habla expresiva que podrían abordarse en el futuro (apartado 6.4).

6.1. Conclusiones generales

El punto de partida de esta tesis ha sido la motivación por avanzar en la mejora de la naturalidad y la expresividad de los sistemas de síntesis del habla. En el *Grup en Processament Multimodal* (GPMM), se partía de la experiencia previa en sistemas de síntesis por concatenación de difonemas y trifonemas (con una única realización de cada unidad) en español y en catalán, orientados a tareas de propósito general (Guaus y Iriondo, 2000; Alías y Iriondo, 2002), que proporcionaban una inteligibilidad suficiente, pero carecían de la naturalidad necesaria para su integración en todo un conjunto emergente de aplicaciones multimedia relacionadas con la interacción persona-máquina. Por otra parte, se desarrolló un sistema de CTH de dominio restringido con el que se obtuvo una calidad muy elevada gracias a un diseño que aprovechaba la gran redundancia y la poca variabilidad del texto de entrada (Alías et al., 2005). Para conseguir una elevada naturalidad y una expresividad adecuada a la aplicación requerida —en este caso, un hombre del tiempo virtual— bastó con un diseño preciso del corpus, una grabación de muy buena calidad, un módulo de selección de unidades que minimiza el número de puntos de concatenación, la recuperación de la prosodia del corpus y un suavizado del contorno de F_0 en la unión entre los segmentos.

Teniendo presente este objetivo de avanzar hacia una síntesis del habla expresiva, se inicia la investigación abordada en la presente tesis, centrada, en un primer momento, en aquellos aspectos en los que los recursos disponibles eran insuficientes. Se constató la necesidad de desarrollar un nuevo corpus oral expresivo y de investigar sobre el modelado y la estimación de la prosodia aplicados al habla expresiva. Los logros conseguidos en estos dos primeros ámbitos suponen un avance hacia el objetivo final, pero no son suficientes para dar por finalizada la investigación en torno a la síntesis del habla expresiva. La experiencia y el conocimiento adquiridos, junto con el trabajo de otros miembros del GPMM, nos permiten sentar las bases de esta línea de investigación que, seguramente, proporcionará desarrollos concretos que tendrán cabida en nuevas aplicaciones.

Algunos aspectos de la investigación realizada han estado condicionados por ciertas limitaciones temporales o de recursos. Por ejemplo, en un sentido más amplio de la síntesis del habla expresiva, sería necesario un abanico emocional mucho más amplio que el cubierto en la presente tesis. Cabe destacar que, en una primera aproximación a la síntesis del habla expresiva, hemos preferido partir de una propuesta basada en un conjunto limitado de estilos, de forma que se acotaba la problemática de la cobertura segmental y prosódica. En esta línea, se tendría que replantear el modelo emocional y la obtención de la correspondiente habla emocionada. En primer lugar, nos debemos cuestionar si se continua con un modelo discreto de emociones. En caso afirmativo, sería preciso evaluar qué emociones serían las más útiles y, si no es el caso, cabría considerar un modelo dimensional (Schröder, 2004). Esta decisión debería tener en cuenta también qué modelo representa mejor la expresión facial de emociones, ya que ambas modalidades, habla e imagen, están relacionados en numerosas aplicaciones multimedia.

Otra decisión importante se relaciona con la metodología seguida para conseguir habla emocionada o expresiva. En el presente trabajo, se ha optado por la grabación en estudio de una locutora profesional que ha interpretado un conjunto de frases que facilitaban el estilo deseado. En diferentes fases del diseño y la producción del corpus se ha contado con la ayuda de los expertos en comunicación audiovisual del LAICOM-UAB. Su colaboración ha sido decisiva en la definición de los cinco estilos expresivos: neutro, sensual, alegre, agresivo y triste; han proporcionado el corpus textual de publicidad que ha servido de base para la definición de las frases que se han grabado; y, finalmente, han colaborado durante la grabación del corpus mediante la aportación de una locutora que es miembro del LAICOM-UAB y de un experto que ha supervisado la correcta expresión de cada estilo. Debido a que la grabación de un estilo podía requerir varias horas, era necesario mantener una expresividad coherente durante y a lo largo de las diferentes sesiones para garantizar un buen resultado.

Este tipo de estrategia seguida para obtener habla expresiva está ligada a la capacidad del actor o locutor para emular determinados estados de ánimo, emociones o actitudes, de modo que éstos se reflejen en su modo de hablar. Aunque se cuente con un locutor profesional o con mucha experiencia para realizar la grabación, puede ocurrir que ciertos enunciados del corpus no se correspondan al estilo expresivo deseado, bien porque se confundan con otros estilos del corpus, bien porque no presenten la suficiente intensidad expresiva. La presente tesis presenta una aportación innovadora en el ámbito de la validación de la expresividad de un corpus, proponiendo un método en el que se

combinan pruebas subjetivas con métodos de identificación automática de emociones en el habla. En el apartado 6.2 se profundiza en las conclusiones y en las posibles líneas de investigación relacionadas con un aspecto que es fundamental en muchas aplicaciones de las tecnologías del habla: el desarrollo y la validación de los corpus orales, en nuestro caso de habla expresiva.

La consecución del corpus de habla expresiva permitió abordar el trabajo de investigación sobre modelado y estimación de la prosodia. Se optó por una estrategia de aprendizaje artificial basado en corpus para aprovechar al máximo el recurso desarrollado y que, además, podía contrarrestar la escasez de conocimiento experto sobre la fonética del habla expresiva. La aplicación de técnicas de aprendizaje automático relativamente sencillas nos ha permitido, en primer lugar, disponer de un sistema completamente automático para generar rasgos prosódicos a partir de un texto así como reproducir los estilos presentes en el corpus. En segundo lugar, se ha llevado a cabo una evaluación del sistema para cada estilo mediante la utilización de medidas objetivas y la realización de pruebas de percepción con oyentes. En el apartado 6.3, se presentan detalladamente las conclusiones y las líneas de trabajo futuras en este ámbito.

Por último, cabe destacar que en la presente tesis no se ha abordado un estudio profundo de las ventajas e inconvenientes de las técnicas de síntesis del habla existentes ni de su aplicación para generar habla emocionada. Por lo tanto, el habla expresiva sintetizada en el ámbito del presente trabajo se basa en el corpus oral y en el modelado prosódico desarrollados, aunque utiliza una técnica de síntesis concatenativa que todavía está en fase de desarrollo en el seno del GPMM (véase el apartado 5.4.1). El módulo final de síntesis de la señal de voz podrá implementarse siguiendo diferentes estrategias (véase el apartado 2.3.2). En el apartado 6.4 se discuten las futuras líneas de investigación relacionadas con este módulo que se consideran apropiadas para conseguir una mejora importante en el resultado final de la síntesis.

El presente trabajo de tesis ha propiciado numerosas contribuciones a congresos internacionales y nacionales, y se ha enriquecido de la participación de su autor en diferentes proyectos de investigación y desarrollo de ámbito europeo y nacional. En el anexo A se presenta un resumen con las principales aportaciones realizadas.

6.2. El corpus de habla emocionada

Entre la comunidad científica dedicada a la síntesis del habla existe un alto grado de consenso sobre la conveniencia de obtener habla grabada por locutores o actores profesionales a pesar de su posible falta de autenticidad (Cowie et al., 2005). Asumiendo que esta posible limitación queda compensada por la calidad de la señal que se ha obtenido al realizar la grabación en un estudio profesional, se ha decidido producir un corpus oral siguiendo esta estrategia con la finalidad de poder avanzar en los diferentes procesos que intervienen en la síntesis del habla expresiva. Bajo esta premisa, una vez grabado el corpus, en el que se recogen cinco estilos diferentes, se hace indispensable una validación de la expresividad del mismo. Por un lado, la evaluación subjetiva es el mejor método para

este propósito. Sin embargo, la evaluación exhaustiva de todas las frases del corpus sería excesivamente costosa en corpus de gran tamaño, como los que se utilizan habitualmente en la síntesis del habla basada en selección de unidades. Por otro lado, no existe suficiente conocimiento científico para emular completamente la percepción subjetiva mediante técnicas automáticas que permitan una validación exhaustiva y fiable de los corpus orales. En el presente trabajo se ha propuesto un método que supone un avance hacia una solución práctica y eficiente de este problema, mediante la combinación de una evaluación subjetiva con técnicas de identificación automática de la emoción en el habla.

El método presentado proporciona un refinamiento automático de la totalidad de un corpus de habla expresiva. Dicho método se ha aplicado al corpus desarrollado en el ámbito de esta tesis, del cual se han detallado el diseño, la grabación y el etiquetado llevados a cabo. Inicialmente, se definió una validación objetiva inspirada en experimentos previos de identificación de emociones en el habla, realizados a partir de la aplicación de técnicas de aprendizaje automático (clasificación). Los porcentajes de clasificación correcta fueron tan elevados que únicamente nos permitieron concluir que los cinco estilos eran acústicamente lo bastante diferentes para que un sistema automático fuera capaz de discriminarlos. También, se llevó a cabo una prueba auditiva con una pequeña parte de los enunciados del corpus, en la que los participantes tenían que identificar las emociones de los estímulos presentados. Los oyentes mostraron mayor confusión que el sistema automático y, especialmente, entre estilos que el sistema no confundía en la misma proporción. Entonces surgió la idea de incluir los resultados de la prueba subjetiva en el sistema automático, con el objetivo de emular el criterio subjetivo característico de la percepción de la emoción en el habla.

En el entrenamiento del sistema automático se incorporó una selección de atributos que tratase de acercar los resultados de la clasificación automática a los de la evaluación subjetiva realizada sobre una pequeña parte del corpus. El método se ha probado con diferentes clasificadores y para cada uno se ha obtenido el subconjunto de atributos que conseguía mejores resultados. Finalmente, mediante una técnica de combinación de clasificadores —*stacking*— se ha refinado el corpus entero, obteniéndose una lista de frases confusas desde el punto de vista expresivo. Con el fin de validar estos resultados, se ha realizado una segunda prueba subjetiva que nos ha permitido comprobar que existe un alto grado de correspondencia entre las decisiones automáticas del sistema desarrollado y la percepción subjetiva mostrada por los participantes de la primera prueba de escucha.

La metodología seguida y los resultados obtenidos en todo el proceso de producción y validación del corpus han permitido cerrar un ciclo completo. Esto no significa que se dé por finalizada la investigación en este campo y, por lo tanto, se sugieren las futuras líneas siguientes:

- Respecto a la obtención de habla emocionada, en futuros trabajos se pueden explorar otras vías de obtenerla como, por ejemplo, reutilizando material almacenado del mundo del cine, la televisión o la radio. Esta opción requiere el desarrollo de herramientas de segmentación automática y de etiquetado que faciliten la gestión de dicho material. También hay que solucionar problemas legales relacionados con la propiedad y el uso de este tipo de audio. Las ventajas de este planteamiento residen

en la eliminación del coste de nuevas grabaciones, un nivel de autenticidad emocional suficiente y una calidad de la señal adecuada.

- En lo que se refiere al sistema de revisión automática, se podrían introducir mejoras en los diferentes módulos del método propuesto: *i)* el alcance de la prueba subjetiva previa y las reglas para determinar el nivel de expresividad de los enunciados evaluados según las respuestas de los oyentes; y *ii)* el ajuste del sistema automático en sus diferentes componentes (la parametrización acústica, el método de selección de atributos y la combinación de nuevos algoritmos de clasificación).

6.3. Modelado de la prosodia basado en corpus

La síntesis del habla expresiva de alta calidad requiere un modelado preciso de los diferentes parámetros acústicos que intervienen en la transmisión de un estado de ánimo o de una intención concreta a través del habla. En esta tesis se ha partido de los parámetros más utilizados en la cuantificación de la prosodia: la curva de F_0, la energía y la duración segmental del habla. Estos parámetros están relacionados con los fenómenos lingüísticos de ámbito suprasegmental: la melodía y la entonación, el ritmo y el acento; por tanto, intervienen también las funciones lingüística, paralingüística y extralingüística de la prosodia. Se ha presentado un método basado en corpus de estimación de estos parámetros prosódicos a partir del texto y se ha probado con los cinco estilos expresivos que componen el corpus desarrollado. Este método permite reproducir unos patrones aprendidos que reflejan las funciones lingüística y paralingüística de la prosodia. La evaluación objetiva nos ha permitido obtener resultados para cada parámetro por separado en cada estilo, comparando la prosodia estimada con la prosodia natural extraída de las frases interpretadas por la locutora. En cambio, en la evaluación subjetiva se ha realizado una valoración global de la prosodia generada automáticamente y, como referencia, se han resintetizado los mismos enunciados pero con la prosodia natural. El análisis de los resultados revela una cierta relación entre la medida del RMSE de la F_0 y la valoración de los oyentes, ya que los estilos ordenados de mayor a menor puntuación DMOS (véase la figura 5.30) prácticamente coinciden con el orden de menor a mayor valor de RMSE de la F_0 (véase la tabla 5.20). En cambio, no se encuentran otras relaciones directas con el resto de medidas. De todas formas, este paralelismo entre resultados subjetivos y una única medida objetiva abre la posibilidad de plantear nuevas hipótesis:

- De los tres rasgos prosódicos estudiados, la F_0 es el parámetro prosódico más relevante en la transmisión de un estilo expresivo. Un hecho que favorecería la validez de esta hipótesis es que se trata del parámetro acústico más estudiado en el habla emocional.

- El contorno de F_0 es más difícil de modelar que la intensidad o la duración segmental y, por tanto, la exactitud del modelo usado puede tener una mayor influencia en la percepción subjetiva.

Los estímulos utilizados en la prueba subjetiva se han escogido del conjunto de test de forma que se ha cubierto un amplio rango de valores de RMSE de la F_0. Los quince enunciados escogidos para cada estilo se pueden dividir en tres grupos que presentan valores de RMSE de F_0 cercanos a los tres cuartiles. El análisis de los resultados (véase la tabla 5.22) no ha permitido confirmar la hipótesis de que los enunciados con RMSE cercano al primer cuartil tendrían mejor puntuación que el resto. Por lo tanto, se ha encontrado una cierta relación entre el RMSE de la F_0 y la percepción de los oyentes en lo que se refiere al estilo, pero esta relación no parece darse si se consideran individualmente las frases que reflejan un mismo estilo.

Los resultados obtenidos al comparar las parejas de estímulos con prosodia natural y prosodia sintética son esperanzadores, ya que muestran un alto grado de parecido entre ambos; incluso se se han dado casos esporádicos de notas más altas para los estímulos con prosodia sintética que natural. Además, se debe tener en cuenta un efecto no deseable presente en la generación de los estímulos de la prueba subjetiva: la resíntesis mediante un método basado en TD-PSOLA tiene el inconveniente de distorsionar más la señal de voz a medida que aumentan las modificaciones de la F_0 y de la duración segmental. Por lo tanto, los estímulos con prosodia natural, a pesar de estar resintetizados, también ganaban en calidad segmental de la señal de voz. Aunque se indique a los oyentes que se centren en la prosodia, es muy difícil que separen ambas componentes.

En el modelado de la prosodia propuesto se pueden introducir numerosas mejoras que van desde el análisis del texto a la evaluación final. En concreto se perfilan las futuras líneas siguientes:

- La definición de la unidad mínima para el modelado de la melodía —el grupo acentual, que coincide con el grupo tónico descrito en Garrido (2001)— ha estado condicionada por una decisión técnica relacionada con la segmentación del texto en este tipo de unidades. Sin embargo, se debería explorar si los resultados del sistema mejoran utilizando una definición de grupo acentual basada en la sílaba tónica y no en la palabra acentuada.
- El análisis del texto se debería enriquecer para generar nuevos atributos prosódicos que permitiesen una mejor descripción del contexto y de la naturaleza de la unidad sobre la cual se lleva a cabo la estimación de un determinado rasgo prosódico. En el caso de la curva de F_0, sería interesante incorporar un análisis sintáctico del texto.
- Se podría avanzar en la representación del contorno de F_0 mediante la inclusión de más valores de F_0 para su aproximación, así como mediante el estudio de otro tipo de funciones diferente a los polinomios para representar de forma paramétriada una determinada curva.
- Con el método de aprendizaje utilizado —el CBR— se han conseguido resultados comparables a los que se pueden obtener con los métodos mayoritariamente utilizados en la comunidad científica —ANN o CART— sin haberse explorado toda su potencialidad. Por lo tanto, en un futuro se puede estudiar la mejora de las distintas fases que componen el CBR, tanto en el entrenamiento como en la explotación del sistema.

- La adaptación a otros idiomas requiere numerosos ajustes o cambios relacionados con el análisis del texto, con los atributos prosódicos extraídos a partir de este análisis y con la definición de las unidades básicas de la prosodia. Como primer paso se debería distinguir entre lenguas con tendencia al compás silábico —p.ej. el español, el catalán o el francés— y lenguas que tienden a un compás acentual —p.ej. el alemán o el inglés— (Ríos, 1991) y profundizar en el análisis de sus propiedades rítmicas.

- El estudio de la relevancia que tiene el modelado de cada rasgo prosódico en la calidad del habla sintetizada para cada estilo expresivo puede resultar muy importante para conocer hasta qué punto la estimación que se obtiene es suficiente o necesita un mayor grado de precisión. La evaluación subjetiva realizada se ha llevado a cabo de forma conjunta, es decir, considerando simultáneamente los tres parámetros prosódicos estimados. Una evaluación subjetiva más completa en la que se preparasen estímulos con un único rasgo prosódico estimado podría aportar nuevos datos sobre la relevancia de cada rasgo prosódico para una emoción determinada (Mixdorff y Jokisch, 2003).

- Finalmente, en relación con los métodos de evaluación de la prosodia estimada, se debería avanzar en el estudio de técnicas objetivas o automáticas que reflejasen el criterio subjetivo. De esta forma, el ajuste del sistema de predicción de la prosodia mediante estas nuevas métricas se podría traducir directamente en una mejora también desde el punto de vista de la percepción subjetiva.

6.4. Síntesis del habla expresiva

El trabajo futuro en este campo está dirigido principalmente a la consecución de un sistema de conversión de texto en habla expresiva en el cual la entrada especifique la emoción requerida. La síntesis del habla expresiva es más exigente en cuanto a la versatilidad del módulo de síntesis de voz. Además del control de la prosodia, la modificación de parámetros relacionados con la cualidad de la voz es un reto para futuros trabajos. A partir del trabajo desarrollado, surgen nuevas preguntas que abren posibles líneas de investigación que les puedan dar respuesta:

- La obtención de nuevos modelos prosódicos a partir de la combinación de los que se han desarrollado, ¿daría lugar a estilos o emociones intermedias con una calidad aceptable? Una representación dimensional de las emociones (Schröder, 2004) y la correspondiente ubicación de los estilos presentes en el corpus desarrollado podría ayudar a extraer ciertas reglas que facilitaran esta combinación.

- La síntesis por selección de unidades trata de recuperar del corpus aquellas unidades que minimizan un función de coste basada normalmente en atributos prosódicos. ¿La organización de un corpus multi-estilo debe ser *blending* o *tiering* (Black, 2003)? Una vez seleccionada la secuencia de unidades óptima, ¿se debe modificar la señal para adaptarse a los valores requeridos? o ¿es mejor concatenar las unidades directamente para evitar la distorsión inherente a la transformación de la señal?

- En relación con el punto anterior, si se sigue la línea de modificar la señal, otros métodos diferentes a TD-PSOLA como, por ejemplo, los basados en modelos sinusoidales más ruido (Stylianou, 2001; Iriondo et al., 2003), ¿conseguirán mejores resultados para este tipo de habla, ya que se podrá tener también un control de los parámetros relacionados con la cualidad de la voz?
- Para este tipo de habla, un cambio de estrategia de síntesis como puede ser la basada en HMM (Yamagishi et al., 2003; Tachibana et al., 2004), que consigue una calidad más uniforme aunque a costa de cierta pérdida de naturalidad, ¿tendrá mayor aceptación por parte de los oyentes?

Estas líneas de investigación son una realidad incipiente en el seno del GPMM y el autor de la presente tesis ha participado ya en algunas de ellas como, por ejemplo, en el estudio de los parámetros de cualidad de la voz en el habla expresiva (Monzo et al., 2007), la síntesis del habla basada en HMM (Gonzalvo et al., 2007) y la mejora de la síntesis por selección de unidades (Alías et al., 2004a).

En conclusión, además de la aplicación directa de los avances descritos en la presente tesis a la síntesis del habla expresiva, parte del conocimiento adquirido y de los recursos desarrollados pueden tener una aplicación en diferentes áreas de las tecnologías del habla como, por ejemplo, en la identificación de locutor, la transformación de voz, la identificación de emociones en el habla o la validación de corpus orales en general.

Bibliografía

Aamodt, A. y Plaza, E. (1994). "Case-based reasoning: foundational issues, methodological variations, and system approaches". *Artificial Intelligence Communications*, **7(1)**, pp. 39–59.

Agüero, P. D., Wimmer, K. y Bonafonte, A. (2004). "Automatic Analysis and Synthesis of Fujisaki's Intonation Model for TTS". En: *Speech Prosody 2004*, pp. 427–430. Nara, Japan.

Alías, F. y Iriondo, I. (2002). "La evolución de la Síntesis del Habla en Ingeniería La Salle". En: *II Jornadas en Tecnología del Habla*, Granada, España.

Alías, F., Iriondo, I., Formiga, Ll., Gonzalvo, X., Monzo, C. y Sevillano, X. (2005). "High quality Spanish restricted-domain TTS oriented to a weather forecast application". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 2573–2576. Lisbon, Portugal.

Alías, F., Llorà, X., Iriondo, I., Sevillano, X., Formiga, L. y Socoró, J. C. (2004a). "Perception-Guided and Phonetic Clustering Weight Tuning Based on Diphone Pairs for Unit Selection TTS". En: *The 8th International Conference on Spoken Language Processing (Interspeech'2004)*, pp. 1221–1224. Jeju Island, Korea.

Alías, F., Monzo, C. y Socoró, J. C. (2006). "A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming". En: *InterSpeech2006 -International Conference on Spoken Language Processing (ICSLP)*, pp. 1698–1701. Pittsburgh, PA, USA.

Alías, F., Sevillano, X., Barnola, P., Formiga, L., Iriondo, I. y Socoró, J. C. (2004b). "Multidomain Text-to-Speech Conversion". En: *III Jornadas en Tecnología del Habla*, Valencia, España.

Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Màrquez, L., Martí, M. A., Padró, L., Placer, R., Rodríguez, H., Taulé, M. y Turmo, J. (1998). "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text". En: *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada, España.

Averill, J. R. (1980). "A constructivist view of emotion". *Emotion: Theory, research and experience*, **1**, pp. 305–339.

Bagshaw, P. (1998). "Unsupervised training of phone duration and energy models for text-to-speech synthesis". En: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, volumen 2, pp. 17–20. Sidney, Australia.

Bailly, G., Bérar, M., Elisei, F. y Odisio, M. (2003). "Audiovisual speech synthesis". *International Journal of Speech Technology*, **(6)**, pp. 331–346.

Bartneck, C. (2000). *Affective Expressions of Machines*. Proyecto Final de Carrera, Stan Ackerman Institute, Eindhoven.
`http://www.bartneck.de/work/aem.pdf`

Black, A. W. (2002). "Perfect Synthesis for all of the people all of the time". En: *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pp. 167–170. Santa Monica, CA, USA.

Black, A. W. (2003). "Unit Selection and Emotional Speech". En: *The 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, volumen 3, pp. 1649–1652. Geneva, Switzerland.

Black, A. W., Zen, H. y Tokuda, K. (2007). "Statistical Parametric Speech Synthesis". En: *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volumen 4, pp. 1229–1232. Honolulu, USA.

Black, A.W. y Taylor, P. (1994). "CHATR: a generic speech synthesis system". En: *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, volumen II, pp. 983–986. Kyoto, Japan.

Blecua, B. y Acín, V. (1995). "Propuesta de un modelo de intensidad vocálica del castellano y el catalán aplicable a un sistema de conversión de texto a habla". *Procesamiento del Lenguaje Natural*, **17**, pp. 257–271.

Bonafonte, A., Escudero, D. y Riera, M. (2006). "La conversión de texto en habla". En: J. Llisterri y M. J. Machuca (Eds.), *Los sistemas de diálogo*, pp. 177–208. Universitat Autònoma de Barcelona, Servei de Publicacions - Fundación Duques de Soria, Bellaterra - Soria.

Bozkurt, B., Ozturk, O. y Dutoit, T. (2003). "Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection". En: *The 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 277–280. Geneva, Switzerland.

Breen, A. y Jackson, P. (1998). "Non-uniform unit selection and the similarity metric within BT's LAUREATE TTS system". En: *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 201–206. Jenolan Caves, Australia.

Brinckmann, C. y Trouvain, J. (2003). "The Role of Duration Models and Symbolic Representation for Timing in Synthetic Speech". *International Journal of Speech Technology*, **6**, pp. 21–31.

Bulut, M., Narayanan, S. S. y Syrdal, A. K. (2002). "Expressive speech synthesis using a concatenative synthesizer". En: *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pp. 1265–1268. Denver, CO, USA.

Burges, C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery*, **2(2)**, pp. 121–167.

Burkhardt, F. y Sendlmeier, W. F. (2000). "Verification of acoustical correlates of emotional speech using formant-synthesis". En: *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 151–156. Newcastle, Northern Ireland, UK.

Cahn, J. E. (1989). *Generating Expression in Synthesized Speech.* Proyecto Final de Carrera, Massachusetts Institute of Technology.

Campbell, N. W. (1990). "Analog I/O nets for syllable timing". *Speech Communication*, **9**, pp. 56–61.

Campbell, N. W. (2000). "Databases of emotional speech". En: *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 34–38. Newcastle, Northern Ireland, UK.

Campbell, N. W. (2002). "Recording techniques for capturing natural everyday speech". En: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, España.

Campbell, N. W. (2004). "Speech and Expression; the Value of a Longitudinal Corpus". En: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Campbell, N. W. (2005). "Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech". *IEICE - Transactions on Information and Systems*, **E88-D(3)**, pp. 376–383.

Campbell, N. W. (2007). "Evaluation of Text and Speech Systems". volumen 37 de *Text, Speech and Language Technology*, pp. 29–64. Springer, Dordrecht.

Campbell, N. W., Hamza, W., Höge, H., Tao, J. y Bailly, G. (2006). "Editorial of the Special Section on Expressive Speech Synthesis". *IEEE Transactions on Speech and Audio Processing*, **14(4)**.

Campillo, F. y Rodríguez, E. (2006). "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems". *Speech Communication*, **48(8)**, pp. 941–956.

Carreras, X., Chao, I., Padró, L. y Padró, M. (2004). "FreeLing: An Open-Source Suite of Language Analyzers". En: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Córdoba, R., Montero, J. M., Gutiérrez, J. M., Vallejo, J. A., Enríquez, E. y Pardo, J. M. (2002). "Selection of the most significant parameters for duration modelling in a Spanish text-to-speech system using neural networks". *Computer Speech & Language*, **16(2)**, pp. 183–203.

Córdoba, R., Vallejo, J. A., Montero, J. M., Gutiérrez-Arriola, J. M., López, M. A. y Pardo, J. M. (1999). "Automatic modeling of duration in a Spanish text-to-speech system using neural networks". En: *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, pp. 1619–1622. Budapest, Hungary.

Cornelius, R. R. (2000). "Theoretical Approaches to Emotion". En: *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pp. 3–10. Newcastle, Northern Ireland, UK.

Cowie, R. y Cornelius, R. R. (2003). "Describing the emotional states that are expressed in speech". *Speech Communication*, **40**, pp. 5–32.

Cowie, R., Douglas-Cowie, E. y Cox, C. (2005). "Beyond emotion archetypes: databases for emotion modelling using neural networks". *Neural Networks*, **18**, pp. 371–388.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. y Schröder, M. (2000a). "FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time". En: *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pp. 19–24. Newcastle, Northern Ireland, UK.

Cowie, R., Douglas-Cowie, E. y Schröder, M. (Eds.) (2000b). *Speech and Emotion: A Conceptual Framework for Research*. ISCA Tutorial and Research Workshop (ITRW), Newcastle, Northern Ireland, UK.
`http://www.isca-speech.org/archive/speech_emotion`

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. y Taylor, J. G. (2001). "Emotion Recognition in Human Computer Interaction". *IEEE Signal Processing*, **18(1)**, pp. 33–80.

Devillers, L., Vidrascu, L. y Lamel, L. (2005). "Challenges in real-life emotion annotation and machine learning based detection". *Neural Networks*, **18**, pp. 407–422.

Devore, J. L. (2005). *Probabilidad y estadística para ingeniería y ciencias*. Thomson International, Mexico, D.F., 6ª edición.

Douglas-Cowie, E., Campbell, N., Cowie, R. y Roach, P. (2003). "Emotional speech: towards a new generation of databases". *Speech Communication*, **40**, pp. 33–60.

Drioli, C., Tisato, G., Cosi, P. y Tesser, F. (2003). "Emotions and voice quality: experiments with sinusoidal modeling". En: *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03), ISCA Tutorial and Research Workshop*, pp. 127–132. Geneva, Switzerland.

Duda, R. O., Hart, P. E. y Stork, D. G. (2001). *Pattern Classification*. Wiley & Sons, Inc., New York, 2ª edición.

Dutoit, T. (1994). "High quality text-to-speech synthesis: a comparison of four candidate algorithms". En: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'94)*, volumen 1, pp. 565–568. Adelaide, South Australia.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer, Dordrecht.

Eide, E., Aaron, A., Bakis, R., Cohen, P., Donovan, R., Hamza, W., Mathes, T., Picheny, M., Polkosky, M., Smith, M. y Viswanathan, M (2003). "Recent Improvements to the IBM Trainable Speech Synthesis System". En: *In proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 708–711. Hong Kong.

Ekman, P. (1999). "Basic Emotions". En: T. Dalgleish y M. Power (Eds.), *Handbook of Cognition and Emotion*, Wiley & Sons, Ltd., Sussex.

Escudero, D. (2003). *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversion Texto-Voz en Español*. Tesis doctoral, Universidad de Valladolid.

Escudero, D. y Cardeñoso, V. (2007). "Applying data mining techniques to corpus based prosodic modeling". *Speech Communication*, **49(3)**, pp. 213–229.

Escudero, D., Cardeñoso, V. y Bonafonte, A. (2003). "Experimental evaluation of the relevance of prosodic features in Spanish using machine learning techniques". En: *The 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2309–2312. Geneva, Switzerland.

Escudero, D., González, C. y Cardeñoso, V. (2002). ""Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in Spanish"". En: *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pp. 1165–1168. Denver, Colorado, USA.

Farrokhi, A., Ghaemmaghami, S. y Sheikhan, M. (2004). "Estimation of prosodic information for Persian text-to-speech system using a recurrent neural network". En: *Speech Prosody 2004*, pp. 475–478. Nara, Japan.

Febrer, A., Padrell, J. y Bonafonte, A. (1998a). "Modeling Phone Duration: Application to Catalan TTS". En: *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 43–46. Jenolan Caves, Australia.

Febrer, M., Febrer, A., Bonafonte, A. y Esquerra, I. (1998b). "Aneto: a Tool for Prosody Analysis of Speech". En: *First COST-G6 Workshop on Digital Audio Effects (DAFX98)*, pp. 19–21. Barcelona, España.

Frank, E. y Witten, I. H. (1998). "Generating accurate rule sets without global optimization". En: *Proceedings of the 15th International Conference on Machine Learning*, pp. 144–151. Morgan Kaufmann, San Francisco, CA.

François, H. y Boëffard, O. (2002). "The greedy algorithm and its application to the construction of a continuous speech database". En: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, volumen 5, pp. 1420–1426. Las Palmas de Gran Canaria, España.

Fujisaki, H., Ohno, S., Nakamura, K., Guirao, M. y Gurlekian, J. (1994). "Analysis of accent and intonation in Spanish based on a quantitative model". En: *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)*, pp. 355–358. Yokohama, Japan.

Garrido, J. M. (1991). *Modelización de patrones melódicos del español para la síntesis y el reconocimiento del habla*. Universitat Autònoma de Barcelona, Bellaterra.

Garrido, J. M. (1996). *Modelling Spanish Intonation for Text-to-Speech Applications*. Tesis doctoral, Departament de Filologia Espanyola. Facultat de Lletres. Universitat Autònoma de Barcelona.

Garrido, J. M. (2001). "La estructura de las curvas melódicas del español: propuesta de modelización". *Lingüística Española Actual*, **23(2)**, pp. 173–209.

Goldberg, D. E. (1989). *Genetic Algorithms in Search Optimization and Machine Learning*. Addison Wesley, Reading, MA.

Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F. y Monzo, C. (2007). "Mixing HMM-Based Spanish Speech Synthesis with a CBR for Prosody Estimation". En: *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007*, volumen 4885 de *Lecture Notes in Computer Science*, pp. 75–85. Springer, Heidelberg.

Grachten, M. (2006). *Expressivity-Aware Tempo Transformations of Music Performances Using Case Based Reasoning*. Tesis doctoral, Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra.

Guaus, R. y Iriondo, I. (2000). "Diphone-Based Unit Selection for Catalan Text-to-Speech Synthesis". En: *Text, Speech and Dialogue. Third International Workshop, TSD 2000 Brno, Czech Republic, September 13-16, 2000 Proceedings*, volumen 1902 de *Lecture Notes in Computer Science*, pp. 277–282. Springer, Heidelberg.

Hirst, D.J., Ide, N. y Veronis, J. (1994). "Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project". En: *Conference Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, pp. 77–80.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A. y Nogueiras, A. (2002). "Interface databases: Design and collection of a multilingual emotional speech database". En: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, España.

Iida, A., Campbell, N., Higuchi, F. y Yasumura, M. (2003). "A corpus-based speech synthesis system with emotion". *Speech Communication*, **40**, pp. 161–187.

Iida, A., Campbell, N., Iga, S., Higuchi, F. y Yasumura, M. (2000). "A speech synthesis system with emotion for assisting communication". En: *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 167–172. Newcastle, Northern Ireland, UK.

Iriondo, I., Alías, F. y Melenchón, J. (2002). "Un modelo híbrido orientado a la síntesis multimodal del habla". *Procesamiento del Lenguaje Natural*, **29**, pp. 159–163.

Iriondo, I., Alías, F., Melenchón, J. y Llorca, M. A. (2004). "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis". En: *Affective Dialogue Systems. Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings*, volumen 3068 de *Lecture Notes in Computer Science*, pp. 197–208. Springer, Heidelberg.

Iriondo, I., Alías, F., Sanchis, J. y Melenchón, J. (2003). "Hybrid Method Oriented to Concatenative Text-to-Speech Synthesis". En: *The 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, volumen 4, pp. 2953–2956. Geneva, Switzerland.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., Tena, D. y Longhi, L. (2000). "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques". En: *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 161–166. Newcastle, Northern Ireland, UK.

Iriondo, I., Martí, J., Oliver, J., Guaus, R. y Moure, H. (1999). "Hacia una síntesis concatenativa de alta calidad para aplicaciones de conversión texto-habla". *Procesamiento del Lenguaje Natural*, **25**, pp. 109–113.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C. y Martínez, E. (2007a). "Validation of an Expressive Speech Corpus by Mapping Automatic Classification to Subjective Evaluation". En: *Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings*, volumen 4507 de *Lecture Notes in Computer Science*, pp. 646–653. Springer, Heidelberg.

Iriondo, I., Planet, S., Socoró, J. C. y Alías, F. (2007b). "Objective and Subjective Evaluation of an Expressive Speech Corpus". En: *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007*, volumen 4885 de *Lecture Notes in Computer Science*, pp. 86–94. Springer, Heidelberg.

Iriondo, I., Planet, S., Socoró, J. C., Alías, F., Monzo, C. y Martínez, E. (2007c). "Expressive Speech Corpus Validation by Mapping Subjective Perception to Automatic Classification Based on Prosody and Voice Quality". En: *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS'2007)*, Saarbrüken, Germany.

Iriondo, I., Socoró, J. C., Formiga, Ll., Gonzalvo, X., Alías, F. y Miralles, P. (2006). "Modelado y estimación de la prosodia mediante razonamiento basado en casos". En: *IV Jornadas en Tecnología del Habla*, pp. 183–188. Zaragoza, España.

Iriondo, I., Socoró, J.C. y Alías, F. (2007d). "Prosody modelling of Spanish for expressive speech synthesis". En: *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volumen 4, pp. 821–824. Honolulu, HI, USA.

James, W. (1884). "What is an Emotion?" *Mind*, **9**, pp. 188–205.
`http://psychclassics.yorku.ca/James/emotion.htm`

John, G. y Langley, P. (1995). "Estimating Continuous Distributions in Bayesian Classifiers". En: *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI'95)*, pp. 338–34. Morgan Kaufmann, San Francisco, CA.

Juslin, P.Ñ. y Laukka, P. (2003). "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, **129(5)**, pp. 770–814.

Klatt, D. H. (1979). "Synthesis by rule of segmental durations in english sentences". En: B. Lindblom y S. Öhman (Eds.), *Frontiers of Speech Communication*, pp. 287–299. Academic Press, New York.

Klatt, D. H. (1987). "Review of Text to Speech Conversion for English". *Journal of the Acoustical Society of America*, **82(3)**, pp. 737–793.

Kohavi, R. (1995). "The Power of Decision Tables". En: N. Lavrac y S. Wrobel (Eds.), *Proceedings of the European Conference on Machine Learning*, volumen 914 de *Lecture Notes in Artificial Intelligence*, pp. 174–189. Springer Verlag, Heidelberg.

Krishna, N. S. y Murthy, H. A. (2005). "Duration Modeling of Indian Languages Hindi and Telugu". En: *Proceedings of 5th ISCA Workshop on Speech Synthesis*, pp. 197–202. Pittsburgh, PA, USA.

Lee, J., Kang, D., Kim, S. y Sung, K. (1998). "Energy contour generation for a sentence using a neural network learning method". En: *Proceedings of th 5th International Conference on Spoken Language Processing (ICSLP)*, volumen 5, pp. 1991–1994. Sidney, Australia.

Lee, S., Kim, Y. J. y Oh, Y.H. (2000). "A Vector-Regression Tree for Generating Energy Contours". *IEEE Signal Processing Letters*, **7(8)**, pp. 216–218.

Lee, S. y Oh, Y. H. (1999). "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems". *Speech Communication*, **28(4)**, pp. 283–300.

Lemmetty, S. (1999). *Review of Speech Synthesis Technology.* Proyecto Final de Carrera, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology. {http://www.acoustics.hut.fi/$\sim$slemmett/dippa/index.html}

Llisterri, J., Aguilar, L., Garrido, J. M., Machuca, M. J., Marín, R., de la Mota, C. y Ríos, A. (1999). "Fonética y tecnologías del habla". En: J. Blecua, G. Clavería, C. Sánchez y J. Torruella (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, pp. 449–479. Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio, Barcelona.

Llisterri, J., Carbó, C., Machuca, M. J., de la Mota, C., Riera, M. y Ríos, A. (2004). "La conversión de texto en habla: aspectos lingüísticos". En: M. A. Martí y J. Llisterri (Eds.), *Tecnologías del texto y del habla*, pp. 145–186. Edicions de la Universitat de Barcelona y Fundación Duques de Soria, Barcelona.

Llisterri, J., Fernández, N., Gudayol, F., Poyatos, J. J. y Martí, J. (1993). ""Testing user's acceptance of Ciber232, a text to speech system used by blind persons"". En: *Speech and Language Technology for Disabled Persons. Proceedings of an ESCA Workshop*, pp. 203–206. Stockholm, Sweden.

Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M. y Ríos, A. (2003). "Entonación y tecnologías del habla". En: P. Prieto (Ed.), *Teorías de la entonación*, pp. 209–243. Ariel (Lingüística), Barcelona.

Llisterri, J. y Mariño, J. B. (1993). "Spanish adaptation of SAMPA and automatic phonetic transcription". ESPRIT PROJECT 6819 (SAM-A Speech Technology Assessment in Multilingual Applications).

Martí, J. y Niñerola, D. (1987). "SINCAS: un conversor texto-voz en castellano". *Procesamiento del Lenguaje Natural*, **5**, pp. 111–122.

Martínez Celdrán, E. (1984). *Fonética. Con especial referencia a la lengua castellana*. Teide, Barcelona.

Massaro, D. W., Light, J. y Geraci, K. (Eds.) (2001). *Auditory-Visual Speech Processing (AVSP 2001)*. Aalborg, Denmark.

Mattingly, I. G. (1974). "Speech synthesis for phonetic and phonological models". *Current Trends in Linguistics*, **12**, pp. 2451–2487.

Melenchón, J. (2006). "Síntesis Facial Audiovisual Realista Personalizable". DEA en Tecnologies de la informació i les comunicacions i la seva gestió. Universitat Ramon Llull.

Melenchón, J., Alías, F. y Iriondo, I. (2002). "PREVIS: A Person-specific Realistic Virtual Speaker". En: *IEEE International Conference on Multimedia and Expo (ICME'02)*, Lausanne, Switzerland.

Melenchón, J., De la Torre, F., Iriondo, I., Alías, F., Martínez, E. y Vicent, L. (2003). "Text to visual synthesis with appearance models". En: *IEEE International Conference on Image Processing (ICIP)*, pp. 237–240. Barcelona, España.

Melenchón, J., Iriondo, I. y Meler, L. (2005). "Simultaneous and Causal Appearance Learning and Tracking". *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, **5(3)**, pp. 44–54.

Melenchón, J., Meler, L. y Iriondo, I. (2004). "On-the-fly Training". En: *3rd International Workshop on Articulated Motion and Deformable Objects, AMDO 2004. Palma de Mallorca, Spain*, volumen 3179 de *Lecture Notes in Computer Science*, pp. 146–154. Springer, Heidelberg.

Michaelis, D., Gramss, T. y Strube, H. (1997). "Glottal to noise excitation ratio - a new measure for describing pathological voices". *Acustica / acta acustica*, **83**, pp. 800–806.

Miralles, P. (2005). *Modelat de la prosòdia mitjançant aprenentatge analògic aplicat a la síntesi de la parla*. Proyecto Final de Carrera, PFC d'Enginyeria Superior en Informàtica, Universitat Ramon Llull.

Mixdorff, H. y Jokisch, O. (2003). "Evaluating the Quality of an Integrated Model of German Prosody". *International Journal of Speech Technology*, **6(1)**, pp. 45–55.

Mixdorff, H., Luksaneeyawin, S., Charnvivit, P. y Thubthong, N. (2003). "Modeling Rhythmic Variation in Thai and its Application to Speech Synthesis". En: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'2003)*, pp. 2457–2460. Barcelona, España.

Mixdorff, H., Nguyen, D. T. y Wu, N. T. (2005). "Duration Modeling in a Vietnamese Text-to-Speech System". En: *Proceedings of 10th International Conference on Speech and Computer (SPECOM)*, Patras, Greece.

Möbius, B. (2000). "Corpus-based speech synthesis: methods and challenges". *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, **6(4)**, pp. 87–116.

Möbius, B. y van Santen, J. (1996). "Modelling segmental duration in German TTS synthesis". En: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, pp. 2395–2398. Philadelphia, PA, USA.

Montero, J. M., Córdoba, R., Macías Guarasa, J., San-Segundo, R., Gutiérrez Arriola, J. y Pardo, J. M. (2004). "Parameter Selection for Prosodic Modelling in a Restricted-Domain Spanish Text-to-Speech System". En: *IFMIP 2004 4th International Forum on Multimedia and Image Processing (World Automation Congress 2004 (WAC 2004))*, Sevilla, España.

Montero, J. M., D'Haro, L. F., Córdoba, R., Vallejo, J., Gutiérrez Arriola, J. y Pardo, J. M. (2003). "ANN F0 Modeling for Female-Voice Synthesis in Spanish: Restricted and Non-Restricted Domains". En: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'2003)*, pp. 563–566. Barcelona, España.

Montero, J. M., Gutiérrez Arriola, J., Colás, J., Enríquez, E. y Pardo, J. M. (1999a). "Analysis and modelling of emotional speech in Spanish". En: *Proceedings of 14th International Conference of Phonetic Sciences (ICPhS'99)*, pp. 957–960. San Francisco, USA.

Montero, J. M., Gutiérrez Arriola, J., Colás, J., Macías Guarasa, J., Enríquez, E. y Pardo, J. M. (1999b). "Development of an emotional speech synthesiser in Spanish". En: *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, pp. 2099–2102. Budapest, Hungary.

Montero, J. M., Gutiérrez Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S. y Pardo, J. M. (1998). "Emotional speech synthesis: From speech database to TTS". En: *The 5th International Conference on Spoken Language Processing (ICSLP)*, pp. 923–926. Sydney, Australia.

Montero, J.M. (2003). *Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano*. Tesis doctoral, Universidad Politécnica de Madrid.

Montoya, N. (1999). *El uso de la voz en la publicidad audiovisual dirigida a los niños y su eficacia persuasiva*. Tesis doctoral, Departament de Comunicació Audiovisual i Publicitat, Universitat Autònoma de Barcelona.

Montoya, N. (2000). "La voz en los anuncios y su eficacia persuasiva en los niños". *Zer. Revista de estudios de comunicación*, **8**.
`http://www.ehu.es/zer/`

Monzo, C., Socoró, J. C., Iriondo, I. y Alías, F. (2007). "Discriminating expressive speech styles by voice quality parameterization". En: *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS'2007)*, pp. 2081–2084. Saarbrüken, Germany.

Moreno, A., Armengol, E. y Béjar, J. (1994). *Aprendizaje automático*. Edicions UPC, Barcelona.

Moulines, E. y Charpentier, F. (1990). "Pitch-synchronous waveform processing techniques for TTS synthesis using diphones". *Speech Communication*, **9**, pp. 453–467.

Murray, I. R. y Arnott, J. L. (1993). "Toward the simulation of Emotion in Synthetic Speech: A Review of The Literature of Human Vocal Emotion". *Journal of the Acoustic Society of America*, **93(2)**, pp. 1097–1108.

Murray, I. R. y Arnott, J. L. (1995). "Implementation and Testing of a System for Producing Emotion-by-Rule in Synthetic Speech". *Speech Communication*, **16**, pp. 369–390.

Murray, I. R., Edgington, M., Campion, D. y Lynn, J. (2000). "Rule-Based Emotion Synthesis Using Concatenated Speech". En: *Proceedings of the ISCA Workshop on Emotion and Speech*, pp. 173–177. Newcastle, Northern Ireland, UK.

Navas, E., Hernáez, I., Luengo, I., Sánchez, J. y Saratxaga, I. (2005). "Analysis of the Suitability of Common Corpora for Emotional Speech Modelling in Standard Basque". En: *Text, Speech and Dialogue, 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005, Proceedings.*, volumen 3658 de *Lecture Notes in Computer Science*, pp. 265–272. Springer, Heidelberg.

Navas, E., Hernáez, I. y Sánchez, J. M. (2002). "Modelo de duración para conversión texto a voz en euskera". *Procesamiento del Lenguaje Natural*, **29**, pp. 147–152.

Navas, E., Hernáez, I. y Luengo, I. (2006). "An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS". *IEEE Transactions on Audio, Speech and Language Processing*, **14(4)**, pp. 1117–1127.

Nogueiras, A., Moreno, A., Bonafonte, A. y Mariño, J. B. (2001). "Speech Emotion Recognition Using Hidden Markov Models". En: *Proceedings of The 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2679–2682. Aalborg, Denmark.

Oudeyer, P. Y. (2003). "The production and recognition of emotions in speech: features and algorithms". *Int. Journal of Human Computer Interaction*, **59(1-2)**, pp. 157–183. Special issue on Affective Computing.

Pérez, E. H. (2003). "Frecuencia de fonemas". *eRTH Revista electrónica de Tecnología del Habla*, **(1)**.
`http://www.rthabla.es`

Petajan, E. D. (1984). "Automatic Lipreading to Enhance Speech Recognition". En: *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, pp. 265–272. Atlanta, GA, USA.

Pitrelli, J. F., B., R., Eide, E. M., Fernandez, R., Hamza, W. y Picheny, M. A. (2006). "The IBM expressive text-to-speech synthesis system for American English". *IEEE Transactions on Audio, Speech and Language Processing*, **14(4)**, pp. 1099–1108.

Platt, J. C. (1999). "Fast training of support vector machines using sequential minimal optimization". En: *Advances in kernel methods: Support vector learning*, pp. 185–208. MIT Press, Cambridge.

Plutchik, R. (2001). "The nature of emotions". *American Scientist*, **89(4)**, pp. 344–350.

Puigví, D., Jiménez, D. y Fernández, J. M. (1994). "Parametrización de las pausas ortográficas en castellano. Aplicación a un conversor de texto a habla". *Procesamiento del Lenguaje Natural*, **15**.

Quinlan, J. R. (1992). "Learning with continuous classes". En: *Proceedings of the Australian Joint Conference on Artificial Intelligence*, pp. 343–348. World Scientific, Singapore.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco.

Riedi, M. (1995). "A neural-network-based model of segmental duration for speech synthesis". En: *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 599–602. Madrid, España.

Ríos, A. (1991). "Caracterización acústica del ritmo del castellano". Trabajo de investigación de Tercer Ciclo. Departament de Filologia Espanyola, Universitat Autònoma de Barcelona.

Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J. y Longhi, L. (1999). "Modelización acústica de la expresión emocional en el español". *Procesamiento del Lenguaje Natural*, **25**, pp. 159–166.

Scherer, K. R. (1999). "Appraisal theory". En: T. Dalgleish y M. Power (Eds.), *Handbook of Cognition and Emotion*, pp. 637–663. Wiley & Sons, Ltd., New York.

Scherer, Klaus R. (1986). "Vocal affect expression: a review and a model for future research". *Psichological Bulletin*, **99**, pp. 143–165.

Scherer, Klaus R. (1988). *Facets of Emotion: Recent Research*. Lawrence Erlbaum Associates Publishers, New Jersey.

Schröder, M. (2001). "Emotional Speech Synthesis: A Review". En: *The 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, volumen 1, pp. 561–564. Aalborg, Denmark.

Schröder, M. (2004). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. Tesis doctoral, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M. y Gielen, S. (2001). "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis". En: *The 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, volumen 1, pp. 87–90. Aalborg, Denmark.

Schröder, M., Hunecke, A. y Krstulovic, S. (2006). "OpenMary - open source unit selection as the basis for research on expressive synthesis". En: *Proceedings of Blizzard Challenge Workshop 2006*, Pittsburgh, PA, USA.

Schröder, M. y Trouvain, J. (2003). "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching." *International Journal of Speech Technology*, **6**, pp. 365–377.

Schweitzer, A. y Möbius, B. (2003). "On the structure of internal prosodic models". En: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'2003)*, pp. 1301–1304. Barcelona, España.

Silverman, K., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. y Hirschberg, J. (1992). "ToBI: A standard for labelling English prosody". En: *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP)*, pp. 867–870. Banff, Alberta, Canada.

Sánchez, S. (1997). "SinCat/2. Lenguaje para la conversión grafema-fonema". *Informe técnico*, Enginyeria i Arquitectura La Salle.

Stylianou, Y. (2001). "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis". *IEEE Transactions on Speech and audio Processing*, **9(1)**, pp. 21–29.

Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T. y Kobayashi, T. (2004). "HMM-based speech synthesis with various speaking styles using model interpolation". En: *Speech Prosody 2004*, pp. 413–416. Nara, Japan.

Tatham, M. y Morton, K. (2003). *Expression in Speech: Analysis and Synthesis*. Oxford Linguistics. Oxford University Press, New York.

Taylor, P. (2000). "Analysis and Synthesis of Intonation using the Tilt Model". *Journal of Acoustical Society of America*, **107(3)**, pp. 1697–1714.

Teixeira, J. P. y Freitas, D. (2003). "Evaluation of a Segmental Durations Model for TTS". En: N. Mamede, J. Baptista, I. Trancoso y M.G. Nunes (Eds.), *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, volumen 2721 de *Lecture Notes in Computer Science*, pp. 40–48. Springer, Heidelberg.

Tesser, F., Cosi, P., Drioli, C. y Tisato, G. (2004). "Prosodic data driven modelling of a narrative style in FESTIVAL TTS". En: *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 185–190. Pittsburgh, PA, USA.

Tesser, F., Cosi, P., Drioli, C. y Tisato, G. (2005). "Emotional Festival-Mbrola TTS Synthesis". En: *The 9th European Conference on Speech Communication and Technology (Interspeech)*, pp. 505–508. Lisbon, Portugal.

Theune, M., Meijs, K., Heylen, D. y Ordelman, R. (2006). "Generating expressive speech for storytelling applications". *IEEE Transactions on Audio, Speech and Language Processing*, **14(4)**, pp. 1137–1144.

Toda, T. (2003). *High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*. Tesis doctoral, Enara Institute for Science and Technology.

Trouvain, J., Barry, W. J., Nielsen, C. y Andersen, O. (1998). "Implications of Energy Declinations for Speech Synthesis". En: *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 47–52. Jenolan Caves, Australia.

Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. y Narayanan, S. S. (2004). "Constructing Emotional Speech Synthesizers With Limited Speech Database". En: *The 8th International Conference on Spoken Language Processing (Interspeech'2004)*, pp. 1185–1180. Jeju Island, Korea.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Princetown University, Ditton.

UIT-T (1994). "Recomendación P.800: Método para la evaluacion subjetiva de la calidad vocal de los dispositivos generadores de voz". Sector de Normalización de las Telecomunicaciones de Unión Internacional de Telecomunicaciones.
`http://www.itu.int/rec/T-REC-P.85-199406-I/es`

UIT-T (1996). "Recomendación P.800: Métodos de determinación subjetiva de la calidad de transmisión". Sector de Normalización de las Telecomunicaciones de Unión Internacional de Telecomunicaciones.
`http://www.itu.int/rec/T-REC-P.800-199608-I/es`

Vapnik, V.Ñ. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York.

Ververidis, D. y Kotropoulos, C. (2003). "A State of the Art Review on Emotional Speech Databases". En: *Proceedings of the 1st Richmedia Conference*, 109–119. Lausanne, Switzerland.

Ververidis, D. y Kotropoulos, C. (2006). "Emotional speech recognition: Resources, features, and methods". *Speech Communication*, **48(9)**, pp. 1162–1181.

Vine, D. S. G. y Sahandi, R. (2000). "Synthesising emotional speech by concatenating multiple pitch recorded speech units". En: *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 157–160. Newcastle, Northern Ireland, UK.

Wang, Y y Witten, I. H. (1997). "Induction of model trees for predicting continuous classes". En: *Proceedings of Poster Papers of the European Conference on Machine Learning*, pp. 128–137. University of Economics, Faculty of Informatics and Statistics, Prague.

Wells, J. (1993). "SAMPA: Computer readable phonetic alphabet".
`http://www.phon.ucl.ac.uk/home/sampa/`

Witten, I. H. y Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2ª edición.

Yamagishi, J., Onishi, K., Masuko, T. y Kobayashi, T. (2003). "Modeling of various speaking styles and emotions for HMM-based speech synthesis". En: *The 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2461–2464. Geneva, Switzerland.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. y Kitamura, T. (1999). "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis". En: *The 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2347–2350. Budapest, Hungary.

Apéndice A

Aportaciones

En este anexo se resume la divulgación científica asociada al presente trabajo de tesis y la participación de su autor en proyectos financiados con fondos públicos y privados.

A.1. Publicaciones científicas

El presente trabajo de investigación ha proporcionado diferentes aportaciones de interés para la comunidad científica. De hecho, las principales ideas, métodos y resultados que son fruto de la actividad investigadora aquí presentada se han expuesto en diferentes congresos y publicado en revistas de ámbito nacional e internacional.

El impacto del trabajo de investigación sobre la comunidad científica se puede resumir en las siguientes publicaciones:

Internacionales

1. Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., Tena, D. y Longhi, L. (2000). "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques". En: *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 161–166. Newcastle, Northern Ireland, UK.

2. Iriondo, I., Alías, F., Sanchis, J. y Melenchón, J. (2003). "Hybrid Method Oriented to Concatenative Text-to-Speech Synthesis". En: *The 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, volumen 4, pp. 2953–2956. Geneva, Switzerland.

3. Iriondo, I., Alías, F., Melenchón, J. y Llorca, M. A. (2004). "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis". En: *Affective Dialogue Systems. Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings*, volumen 3068 de *Lecture Notes in Computer Science*, pp. 197–208. Springer, Heidelberg.

4. Iriondo, I., Planet, S., Socoró, J. C. y Alías, F. (2007b). "Objective and Subjective Evaluation of an Expressive Speech Corpus". En: *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007*, volumen 4885 de *Lecture Notes in Computer Science*, pp. 86–94. Springer, Heidelberg.

5. Iriondo, I., Planet, S., Alías, F., Socoró, J. C. y Martínez, E. (2007a). "Validation of an Expressive Speech Corpus by Mapping Automatic Classification to Subjective Evaluation". En: *Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings*, volumen 4507 de *Lecture Notes in Computer Science*, pp. 646–653. Springer, Heidelberg.

6. Iriondo, I., Planet, S., Socoró, J. C., Alías, F., Monzo, C. y Martínez, E. (2007c). "Expressive Speech Corpus Validation by Mapping Subjective Perception to Automatic

Classification Based on Prosody and Voice Quality". En: *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS'2007)*, Saarbrüken, Germany.

7. Iriondo, I., Socoró, J.C. y Alías, F. (2007d). "Prosody modelling of Spanish for expressive speech synthesis". En: *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volumen 4, pp. 821–824. Honolulu, HI, USA.

Nacionales

8. Iriondo, I., Martí, J., Oliver, J., Guaus, R. y Moure, H. (1999). "Hacia una síntesis concatenativa de alta calidad para aplicaciones de conversión texto-habla". *Procesamiento del Lenguaje Natural*, **25**, pp. 109–113.

9. Iriondo, I., Alías, F. y Melenchón, J. (2002). "Un modelo híbrido orientado a la síntesis multimodal del habla". *Procesamiento del Lenguaje Natural*, **29**, pp. 159–163.

10. Iriondo, I., Socoró, J. C., Formiga, Ll., Gonzalvo, X., Alías, F. y Miralles, P. (2006). "Modelado y estimación de la prosodia mediante razonamiento basado en casos". En: *IV Jornadas en Tecnología del Habla*, pp. 183–188. Zaragoza, España.

Las publicaciones 1 y 3 se corresponden con las primeras aportaciones relacionadas con el modelado acústico del habla emocional. Ambas contribuciones se presentaron en dos talleres especializados en el tema del habla y la emocion: *The ISCA Workshop on Speech and Emotion* y *Affective Dialogue Systems, Tutorial and Research Workshop* respectivamente. La participación en dichos talleres proporcionó la base de conocimiento necesaria para concretar el enfoque del presente trabajo de tesis.

Las publicaciones 4, 5 y 6 comprenden la descripción del desarrollo del corpus oral y los diferentes avances sobre la evaluación de su contenido expresivo, que comprenden la práctica totalidad del capitulo 4 de la presente tesis.

Las publicaciones 7 y 10 presentan la propuesta de modelado prosódico del habla expresiva basado en CBR y aportan unos resultados preliminares, tanto de evaluación objetiva como subjetiva.

Finalmente, las publicaciones 2, 8 y 9 están centradas en aspectos concretos de la síntesis del habla orientados hacia la mejora de la naturalidad.

En colaboración con otros miembros del GPMM

A continuación se enumeran algunas de las publicaciones en las que ha colaborado el autor de la presente tesis sin ser el primer autor. De un conjunto más amplio, se han seleccionado aquellas que tienen alguna relación que el trabajo desarrollado.

11. Alías, F., Sevillano, X., Barnola, P., Formiga, L., Iriondo, I. y Socoró, J. C. (2004b). "Multidomain Text-to-Speech Conversion". En: *III Jornadas en Tecnología del Habla*, Valencia, España.

12. Alías, F., Llorà, X., Iriondo, I., Sevillano, X., Formiga, L. y Socoró, J. C. (2004a). "Perception-Guided and Phonetic Clustering Weight Tuning Based on Diphone Pairs for Unit Selection TTS". En: *The 8th International Conference on Spoken Language Processing (Interspeech'2004)*, pp. 1221–1224. Jeju Island, Korea.

13. Alías, F., Iriondo, I., Formiga, Ll., Gonzalvo, X., Monzo, C. y Sevillano, X. (2005). "High quality Spanish restricted-domain TTS oriented to a weather forecast application". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 2573–2576. Lisbon, Portugal.

14. Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F. y Monzo, C. (2007). "Mixing HMM-Based Spanish Speech Synthesis with a CBR for Prosody Estimation". En: *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007*, volumen 4885 de *Lecture Notes in Computer Science*, pp. 75–85. Springer, Heidelberg.

15. Guaus, R. y Iriondo, I. (2000). "Diphone-Based Unit Selection for Catalan Text-to-Speech Synthesis". En: *Text, Speech and Dialogue. Third International Workshop, TSD 2000 Brno, Czech Republic, September 13-16, 2000 Proceedings*, volumen 1902 de *Lecture Notes in Computer Science*, pp. 277–282. Springer, Heidelberg.

16. Melenchón, J., Alías, F. y Iriondo, I. (2002). "PREVIS: A Person-specific Realistic Virtual Speaker". En: *IEEE International Conference on Multimedia and Expo (ICME'02)*, Lausanne, Switzerland.

17. Melenchón, J., De la Torre, F., Iriondo, I., Alías, F., Martínez, E. y Vicent, L. (2003). "Text to visual synthesis with appearance models". En: *IEEE International Conference on Image Processing (ICIP)*, pp. 237–240. Barcelona, España.

18. Melenchón, J., Meler, L. y Iriondo, I. (2004). "On-the-fly Training". En: *3rd International Workshop on Articulated Motion and Deformable Objects, AMDO 2004. Palma de Mallorca, Spain*, volumen 3179 de *Lecture Notes in Computer Science*, pp. 146–154. Springer, Heidelberg.

19. Melenchón, J., Iriondo, I. y Meler, L. (2005). "Simultaneous and Causal Appearance Learning and Tracking". *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, **5(3)**, pp. 44–54.

20. Monzo, C., Socoró, J. C., Iriondo, I. y Alías, F. (2007). "Discriminating expressive speech styles by voice quality parameterization". En: *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS'2007)*, pp. 2081–2084. Saarbrüken, Germany.

Las publicaciones 11, 12, 13 y 15 están relacionadas con la estrategia de síntesis que se ha utilizado para generar los estímulos utilizados en la prueba subjetiva de evaluación

del módulo de estimación de la prosodia. La experiencia adquirida, junto con las nuevas líneas de investigación apuntadas en las publicaciones 14 (síntesis basada en HMM) y 20 (modelado de los parámetros de cualidad de la voz), deben marcar el diseño futuro del módulo de síntesis para mejorar la calidad final.

Las publicaciones 16–19 reflejan la colaboración habida con los miembros del área de Visión por Computador del GPMM con el fin de desarrollar personajes virtuales.

A.2. Proyectos de investigación y desarrollo

A continuación se describen los proyectos en los que el autor de la tesis, como miembro del GPMM, ha participado y que tienen relación con el presente trabajo.

A.2.1. Con financiación pública

SALERO[1]: Semantic AudiovisuaL Entertainment Reusable Objects (FP6/2004/IST/4)

Proyecto en curso financiado por el VI Programa Marco de la Unión Europea en el que participan trece socios entre empresas y centros de investigación que finaliza en diciembre de 2009. Su objetivo es facilitar la creación de nuevos productos multimedia como juegos, películas o programas de televisión haciéndola mejor, más rápida y más barata gracias a la combinación de gráficos por ordenador, tecnología del habla y el lenguaje, web semántica y búsquedas basadas en contenido.

SAVE: Síntesis Audiovisual Expresiva (TEC2006-08043/TCM). Ministerio de Educación y Ciencia.

Proyecto de I+D en curso que finaliza al final del año 2009 con el objetivo de generar cabezas parlantes capaces de transmitir estados de ánimo mientras hablan. De la presente tesis, el estudio de la representación emocional y la experiencia obtenida en el desarrollo del corpus oral son de inmediata aplicación a este proyecto.

IntegraTV4all (FIT-350301-2004-2). Ministerio de Ciencia y Tecnología.

Proyecto de I+D desarrollado por la Ingeniería de *software* TMT Factory junto con la fundación ONCE, las universidades Carlos III y Politécnica de Madrid que tiene como objetivo el desarrollo de servicios adaptados de ocio, información y tele-trabajo a través de la televisión para hoteles. Se dotó a dichos servicios de funcionalidades avanzadas de visión y habla asistida para facilitar la estancia a huéspedes con discapacidades sensoriales. El presente trabajo estuvo relacionado con el desarrollo de una cabeza parlante expresiva de fácil personalización, tarea de la que fue responsable el GPMM.

[1] `http://www.salero.info/`

A.2.2. Contratos con empresas

Módulo sintetizador de voz para aplicación a meteorología[2]. *Corporació Catalana de Ràdio i Televisió* (CCRTV)

Desarrollo del módulo de síntesis de habla de alta calidad en un dominio restringido al sistema de generación automática de previsiones meteorológicas y sincronización con el personaje virtual. Se realizó durante el año 2004.

A.2.3. Participación en eventos

El futuro de los sistemas de diálogo (Soria, 11 de julio de 2005)

Participación como ponente en este curso de Tecnologías Lingüísticas dirigido por el Dr. Joaquim Llisterri y organizado por la Fundación Duques de Soria con dos sesiones tituladas "La síntesis multimodal"

Dia de la Ciència a les Escoles (2005-2007)

Se trata de una actividad realizada durante la *Setmana de la Ciència* en la que de forma totalmente simultánea, unos setenta científicos transiten su experiencia investigadora en escuelas de bachillerato y formación profesional repartidas por toda Catalunya. En tres ocasiones: Solsona (2005), Banyoles (2006) y Sant Adrià del Besòs (2007), el autor de la presente tesis ha tenido ocasión de participar exponiendo temas relacionados con la síntesis de personajes virtuales con emociones.

[2] `http://www.meteosam.com/`

Apéndice B

Descripción fonética del corpus

B.1. Inventario de fonemas y alófonos para la síntesis del español

Las dos tablas presentadas en este apartado recogen el inventario de fonemas y alófonos utilizado para la representación fonética del sintetizador en castellano de EALS-URL (Martí y Niñerola, 1987). Se utiliza una notación basada en SAMPA (Llisterri y Mariño, 1993; Wells, 1993) y modificada en función de algunas decisiones relacionadas con el desarrollo del sistema.

La tabla B.1 muestra los símbolos utilizados para los segmentos vocálicos y semivocálicos[1][2]. Los símbolos vocálicos en mayúsculas son propios de nuestro sistema y los utilizamos para diferenciar las vocales tónicas de las átonas, que se representan con la correspondiente letra en minúscula. Esta diferenciación entre vocales átonas y tónicas mediante símbolos distintos ha facilitado la programación de algunos módulos del sistema, como el transcriptor fonético, la segmentación automática, el modelado prosódico o el selector de unidades para la síntesis del habla.

Tabla B.1: Inventario de vocales y semivocales utilizado en la síntesis del español representado mediante una adaptación de SAMPA.

FONEMAS Y ALÓFONOS VOCÁLICOS				
IPA	**SAMPA**	**Adaptación**	**Descripción**	**Ejemplo**
Vocales				
i	i	i	anterior cerrada (átona)	p**i**sar
e	e	e	anterior media (átona)	c**e**rrar
a	a	a	central abierta (átona)	s**a**ber
o	o	o	posterior media (átona)	c**o**mer
u	u	u	posterior cerrada (átona)	s**u**mar
’i	”i	I	anterior cerrada (tónica)	p**i**co
’e	”e	E	anterior media (tónica)	p**e**ro
’a	”a	A	central abierta (tónica)	v**a**lle
’o	”o	O	posterior media (tónica)	t**o**ro
’u	”u	U	posterior cerrada (tónica)	d**u**ro
Semivocales y semiconsonantes				
j	j	j	anterior palatal (diptongo decreciente)	re**y**
			anterior palatal (diptongo creciente)	p**i**e
w	w	w	posterior labiovelar (diptongo decreciente)	de**u**da
			posterior labiovelar (diptongo creciente)	m**u**y

Por otra parte, para los fonemas y alófonos consonánticos se ha utilizado el inventario mostrado en la tabla B.2, que constituye una adaptación de SAMPA (Llisterri

[1] Tradicionalmente, en la fonética española, se ha distinguido entre semiconsonantes y semivocales (Martínez Celdrán, 1984). Sin embargo, en otras lenguas, ambas se denominan conjuntamente con el término inglés *glide*. En este trabajo las hemos transcrito con el mismo símbolo.

[2] En una futura revisión del inventario debería diferenciarse la consonante fricativa palatal sonora de la semivocal o semiconsonante anterior palatal.

y Mariño, 1993), aunque presenta algunas diferencias respecto a la notación original. En primer lugar, se han evitado aquellos símbolos que necesitan dos caracteres para su representación: es el caso de los fonemas /tS/ y /jj/, sustituidos por /C/ y /j/ respectivamente. Respecto al inventario utilizado por Martí y Niñerola (1987), se han añadido los alófonos [N] y [M]. en cambio, no se han incorporado los alófonos [z] y [dZ] pues, al tratarse de realizaciones condicionadas por el contexto fonético, quedan reflejadas en los difonemas empleados. Por otra parte, su inclusión supondría un aumento considerable del número de difonemas.

Tabla B.2: Inventario de fonemas y alófonos consonánticos utilizado en la síntesis del español representado mediante una adaptación de SAMPA.

FONEMAS Y ALÓFONOS CONSONÁNTICOS				
IPA	**SAMPA**	**Adaptación**	**Descripción**	**Ejemplo**
Oclusivas				
p	p	p	bilabial sorda	**p**adre
b	b	b	bilabial sonora	**v**ino
d	t	t	dental sorda	**t**omo
p	d	d	dental sonora	**d**onde
k	k	k	velar sorda	**c**asa
g	g	g	velar sonora	**g**ata
Fricativas				
f	f	f	labiodental sorda	**f**ácil
θ	T	T	interdental sorda	**c**inco
s	s	s	alveolar sorda	**s**ala
y	jj	j	palatal sonora	**hi**elo
x	x	x	velar sorda	mu**j**er
Aproximantes				
β	B	B	bilabial sonora	la**v**a
ð	D	D	dental sonora	na**d**a
ɣ	G	G	velar sonora	lue**g**o
Africada				
tʃ	tS	C	palatal sorda	mu**ch**o
Nasales				
m	m	m	bilabial	**m**ismo
ɱ		M	labiodental	á**n**fora
n	n	n	alveolar	**n**unca
ɲ	J	J	palatal	a**ñ**o
ŋ	N	N	velar	u**n**gir
Laterales				
l	l	l	alveolar	**l**ejos
ʎ	L	L	palatal	caba**ll**o
Vibrantes				
ɾ	r	r	alveolar simple	pu**r**o
r	rr	R	alveolar múltiple	to**rr**e

B.2. Ejemplos de textos del corpus

A continuación se muestran algunos ejemplos de frases que se han utilizado para la grabación de los cinco estilos expresivos. Algunos signos de puntuación se han añadido después de la grabación para indicar aquellas pausas realizadas por la locutora que no estaban marcadas previamente.

B.2.1. Ejemplos de frases publicitarias en el campo de la automoción

¡Oh cielos! Pisa el embrague a fondo.

Fin de semana sin fin. ¡Oh no!

¡Lo mejor que te puede pasar de nuevo! ¡Lo último en diésel!

¿Ha visto alguna vez chocar a un búho?

¿Actor o espectador?

¿El camino más corto?

¿Qué te duele más?

¿Sueñas con un coche que te permita dominar cualquier situación?

Quien conduce, lo sabe.

Tienes que ser un loco, un loco para intentarlo, y alguien brillante para conseguirlo.

La pieza clave de tu coche no viene de serie para un transporte exigente.

La diferencia. En contra y luego estás tú.

Acostúmbrate a verlo de lejos.

Ahora que el futuro sucedió ayer.

Ahora que sabemos que somos menos listos de lo que pensábamos.

No querrás verte de otra forma.

Ahora ha conseguido que el motor de gasolina consuma poco.

Se ha hecho justicia. Cambio automático.

Crear un automóvil desde cero, está bien.

Pero, es mejor hacerlo desde una gran idea.

Cuando la realidad no es suficiente.

Infinitas posibilidades la evolución de la tecnología.

B.2.2. Ejemplos de frases publicitarias en el ámbito de la educación

¡Increíblemente fascinante!

¡La aventura continua!

¡La diversión, es nuestra historia!

¡Tu profesor de idiomas, a domicilio!

¡Qué fácil es encontrarte bien!

¡Se está montando un buen pollo en todas las librerías!

¿Aprender idiomas?

¿Quién ha dicho que los directores de arte no saben dibujar?

¿Quieres aprender una profesión a tu medida?

¿Te gustaría tener un animal y cuidarlo tu solo?

¿Has leído el horóscopo de la semana?

¿Le gustaría desarrollar su memoria?

Aprender lo que más te gusta, es cuestión de práctica.

Aprenderán jugando, estas vacaciones.

De locura, lo mires por donde lo mires, lo más fácil es que te toque.

Demostrado, los mejores números de la lotería, están por detrás.

Desde mil novecientos cincuenta y seis, el espíritu de una enseñanza de calidad.

Desde niños y niñas, hasta doctores en física.

Detrás de todo gran hombre y toda gran mujer, hay siempre una gran aventura.

Dile a tu jefe que te gusta nadar contra corriente.

Disfrute en su hogar, de lo mejor de la música clásica.

Durante quince días, te guardamos el puesto.

El marido, la esposa, el multimillonario, una proposición indecente.

El mejor método para aprender inglés, divirtiéndose.

El país donde los deseos, se hacen realidad.

El placer de la buena lectura.

B.2.3. Ejemplos de frases publicitarias en el campo de las nuevas tecnologías

¡Piensa y trabaja!

¡Por fin es sábado!

¡Que bueno es el placer solitario!

¡Qué no te falte ni uno!

¡Sácale jugo al mundo digital!

¡La mejor música y mucho más!

¿Se imagina un teléfono con trescientos metros de cable?

¿Su empresa utiliza sus ordenadores sólo para escribir a máquina?

Entonces, ¿por qué utilizar sus teléfonos móviles sólo para hacer llamadas?

¿Su sistema de comunicación, podrá adaptarse al cambio que le exija el futuro?

¿Te conformas con mirar, o prefieres participar?

¿Tiene usted ojo crítico?

¿Cuántas veces quiere que le recuerden a lo largo del año?

¿Cuánto quieres cambiar?

Buscamos las cien mejores ideas tecnológicas.

Bajan otra vez los precios de las llamadas internacionales.

Para que te cueste menos, hablar con los tuyos.

En soluciones informáticas cuente con un buen socio.

Aplicaciones informáticas, adecuadas a las necesidades de cada empresa.

Aunque a veces lo olvido, creo que realmente somos muy parecidos.

Complete ahora las hojas de cálculo, más potentes del mundo.

Celebrar nuestro cien cumpleaños tenía que traer muchos cambios.

Claves de la economía mundial.

Avanzando el arte de imprimir.

Ayudamos, a predecir el tiempo.

Ahora puedes guardar aquí, todo lo que escribes allí.

B.2.4. Ejemplos de frases publicitarias en el ámbito de la cosmética

¡Lo tiene, todo!

¡Enamórate!

¡Fuera el estrés!

¡No tengas sorpresas este verano!

¡Vivan los dos mil!

¿Ardor de estómago?

¿Ampollas y rozaduras?

¿El fin de la barra de labios tradicional?

¿Está preparada para un cuerpo, perfecto?

¿Hasta qué punto aprecias tus miembros?

¿La mejor protección, y la máxima resistencia a la arena y el agua?

¿Lleva hoy su protección antioxidante?

¿Por qué cinco regalos distintos?

¿Qué te vas a poner hoy?

¿Y si se pudiese retrasar el tiempo?

El noventa y seis coma uno por ciento de nuestros clientes, están satisfechos.

Incluso, dentro de treinta días, mi pelo mantendrá la viveza de su color.

A prueba de roces.

Aire de mujer.

Tu piel es incapaz de fabricar sus propios lípidos.

La pintura de labios que lo resiste, todo.

Su cabello, corre peligro.

Cada mañana, toda la energía de la uva para una piel, apetecible, y jugosa.

El fuego de la pasión.

La última seducción.

Olores y colores, me inundan de emoción y de sensaciones.

B.2.5. Ejemplos de frases publicitarias en el ámbito de los viajes

¡Hay que ver Ceuta! Algo estrecho nos une.

¡Benefíciate de precios redondos, en nuestros supervuelos!

¿Por qué ha quemado su dinero, estas vacaciones de semana santa?

¿Conoce usted Portugal?

¿A qué espera, para venir?

¿Dónde reside la clave de la globalización?

¿En cuánto territorio se cubre, o en cómo se cubre?

¿Está seguro que más estrellas significan, mejor servicio?

Todo el mar, en siete metros.

Como en su casa.

El placer de conducir. El placer de viajar.

Disfrute de las ventajas, del invierno.

Disfrútela. Antes de que se ponga de moda.

El contenido de nuestros pabellones.

India, aún más hermosa de lo que imagina.

La isla tropical más sorprendente del mundo.

Mira. Descubrirás que nunca antes habías visto el azul.

Cada día hay más gente, que desayuna con nosotros.

Consiga unas vacaciones de ensueño, mientras vuela.

Por encima de todo.

Usted sube y baja, llega, y se va en un abrir y cerrar de ojos. El tiempo, lo es todo.

Setenta años, trabajando, y creciendo.

Adelante su reserva, saldrá ganando.

A los portugueses, nos gusta cuidar las formas.

B.3. Difonemas y trifonemas del corpus en español

Las tablas de este apartado muestran la lista completa de difonemas y trifonemas incluyendo las correspondientes secuencias portadoras necesarias para su grabación, así como las transcripciones fonéticas de éstas. La creación de estas tablas se ha llevado a cabo en dos etapas claramente diferenciadas. Las primeras 698 unidades (hasta la palabra *Quechua* de la tabla B.8) ya estaban definidas al inicio del presente trabajo fruto del desarrollo del sistema de CTH SINCAS (Martí y Niñerola, 1987). En una segunda etapa, desarrollada en el ámbito del presente trabajo, se ha completado el inventario de difonemas y trifonemas, ya que se ha diferenciado entre vocales tónicas y átonas y se han incluido los alófonos [N] y [M] (véanse las tablas B.3 a B.12).

Tabla B.3: Lista de difonemas y trifonemas (I).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ pi /	pisar	/ _pisAR_ /	/ pE /	espejo	/ _espExo_ /
/ pA /	pala	/ _pAla_ /	/ pO /	poco	/ _pOko_ /
/ pU /	puro	/ _pUro_ /	/ bI /	vino	/ _bIno_ /
/ be /	beber	/ _beBER_ /	/ ba /	barniz	/ _barnIT_ /
/ bO /	bolsa	/ _bOlsa_ /	/ bu /	buscar	/ _buskAR_ /
/ tm /	atmosférico	/ _atmosfEriko_ /	/ tl /	atlético	/ _atlEtiko_ /
/ ti /	timón	/ _timOn_ /	/ te /	tener	/ _tenER_ /
/ tA /	taza	/ _tATa_ /	/ to /	tocino	/ _toTIno_ /
/ tU /	tu	/ _tU_ /	/ di /	directo	/ _dirEkto_ /
/ dE /	dedo	/ _dEDo_ /	/ dA /	dado	/ _dADo_ /
/ do /	dominó	/ _dominO_ /	/ dU /	duda	/ _dUDa_ /
/ kI /	quilo	/ _kIlo_ /	/ kE /	queso	/ _kEso_ /
/ kA /	cara	/ _kAra_ /	/ kO /	codo	/ _kODo_ /
/ kU /	curso	/ _kUrso_ /	/ gi /	guisar	/ _gisAR_ /
/ gE /	guerra	/ _gERa_ /	/ gA /	ganga	/ _gANGa_ /
/ gO /	gorro	/ _gORo_ /	/ gU /	guno	/ _gUno_ /
/ mp /	campo	/ _kAmpo_ /	/ mB /	cambio	/ _kAmBjo_ /
/ mm /	Ammón	/ _ammOn_ /	/ mn /	amnesia	/ _amnEsja_ /
/ mi /	mitad	/ _mitAD_ /	/ mE /	mesa	/ _mEsa_ /
/ mA /	mano	/ _mAno_ /	/ mO /	momia	/ _mOmja_ /
/ mU /	musa	/ _mUsa_ /	/ nt /	un tío	/ _UntIo_ /
/ nD /	donde	/ _dOnDe_ /	/ nm /	un metro	/ _UnmEtro_ /
/ nn /	ennegrecer	/ _enneGreTER_ /	/ nJ /	un ñoño	/ _UnJOJo_ /
/ nT /	encima	/ _enTIma_ /	/ ns /	ensordecer	/ _ensorDeTER_ /
/ nC /	encharcar	/ _enCarkAR_ /	/ nl /	enlace	/ _enlATe_ /
/ nL /	caen llaves	/ _kAenLABes_ /	/ nR /	enredar	/ _enReDAR_ /
/ nI /	nido	/ _nIDo_ /	/ ne /	nevar	/ _neBAR_ /
/ na /	luna	/ _lUna_ /	/ no /	mano	/ _mAno_ /
/ nU /	nube	/ _nUBe_ /	/ n_ /	don	/ _dOn_ /
/ JI /	cañí	/ _kaJI_ /	/ JE /	niñera	/ _niJEra_ /
/ Ja /	niña	/ _nIJa_ /	/ Jo /	niño	/ _nIJo_ /
/ JU /	ñu	/ _JU_ /	/ Nk /	encasillar	/ _eNkasiLAR_ /
/ NG /	enganchar	/ _eNGanCAR_ /	/ Nx /	enjaular	/ _eNxawlAR_ /
/ Mf /	ánfora	/ _AMfora_ /	/ Bp /	Jacob puede	/ _xakOBpwEDe_ /
/ Bt /	Jacob toca	/ _xakOBtOka_ /	/ Bd /	abdomen	/ _aBdOmen_ /
/ Bb /	Jacob bebe	/ _xakOBbEBe_ /	/ Bk /	Jacob cub	/ _xakOBkUB_ /
/ Bg /	Jacob gasta	/ _xakOBgAsta_ /	/ Bm /	Jacob muerde	/ _xakOBmwErDe_ /
/ Bn /	abnegar	/ _aBneGAR_ /	/ BJ /	Jacob ñoño	/ _xakOBJOJo_ /
/ Bf /	Jacob fiero	/ _xakOBfjEro_ /	/ BT /	Jacob cerca	/ _xakOBTErka_ /
/ Bs /	ábside	/ _ABsiDe_ /	/ BC /	Jacob chato	/ _xakOBCAto_ /
/ Bx /	abjurar	/ _aBxurAR_ /	/ BL /	Jacob llave	/ _xakOBLABe_ /
/ BR /	Jacob ruso	/ _xakOBRUso_ /	/ Bj /	Abyecto	/ _aBjEkto_ /
/ BI /	aviso	/ _aBIso_ /	/ BE /	abeja	/ _aBExa_ /
/ Ba /	haba	/ _ABa_ /	/ Bo /	cubo	/ _kUBo_ /
/ Bu /	abusar	/ _aBusAR_ /	/ B_ /	Jacob	/ _xakOB_ /
/ fp /	Calaf playa	/ _kalAfplAja_ /	/ fb /	Calaf busca	/ _kalAfbUska_ /
/ ft /	Calaf tierno	/ _kalAftjErno_ /	/ fd /	Calaf dentro	/ _kalAfdEntro_ /
/ fk /	Calaf cala	/ _kalAfkAla_ /	/ fm /	Calaf menta	/ _kalAfmEnta_ /
/ fn /	Calaf noche	/ _kalAfnOCe_ /	/ fJ /	Calaf ñoño	/ _kalAfJOJo_ /
/ ff /	Calaf familiar	/ _kalAffamiljAR_ /	/ fT /	Calaf cerca	/ _kalAfTErka_ /
/ fs /	Calaf sereno	/ _kalAfserEno_ /	/ fC /	Calaf chato	/ _kalAfCAto_ /
/ fx /	Calaf justo	/ _kalAfxUsto_ /	/ fG /	afgano	/ _afGAno_ /
/ fL /	Calaf lleno	/ _kalAfLEno_ /	/ fR /	Calaf ruin	/ _kalAfRwIn_ /
/ fI /	firma	/ _fIrma_ /	/ fE /	feo	/ _fEo_ /
/ fA /	fama	/ _fAma_ /	/ fO /	forma	/ _fOrma_ /
/ fU /	fusa	/ _fUsa_ /	/ f_ /	Calaf	/ _kalAf_ /
/ Tp /	haz pobre	/ _ATpOBre_ /	/ Tb /	haz bien	/ _ATbjEn_ /
/ Tt /	haz temblar	/ _ATtemBlAR_ /	/ Td /	haz daño	/ _ATdAJo_ /
/ Tk /	haz caso	/ _ATkAso_ /	/ Tg /	haz guerra	/ _ATgERa_ /
/ Tm /	haz miedo	/ _ATmjEDo_ /	/ Tn /	haz nada	/ _ATnADa_ /
/ TJ /	haz ñoño	/ _ATJOJo_ /	/ Tf /	haz faena	/ _ATfaEna_ /
/ TT /	haz zapatos	/ _ATTapAtos_ /	/ Ts /	haz siesta	/ _ATsjEsta_ /
/ TC /	haz chistes	/ _ATCIstes_ /	/ Tx /	haz jota	/ _ATxOta_ /
/ Tl /	haz leña	/ _ATlEJa_ /	/ TL /	haz llaves	/ _ATLABes_ /

Tabla B.4: Lista de difonemas y trifonemas (II).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ TR /	haz reloj	/ _ATRelOx_ /	/ TI /	cine	/ _TIne_ /
/ TE /	cesta	/ _TEsta_ /	/ Ta /	zapato	/ _TapAto_ /
/ To /	zoquete	/ _TokEte_ /	/ TU /	zumo	/ _TUmo_ /
/ T_ /	haz	/ _AT_ /	/ Dp /	alud puro	/ _alUDpUro_ /
/ Db /	alud bueno	/ _alUDbwEno_ /	/ Dt /	alud total	/ _alUDtotAl_ /
/ Dd /	alud duro	/ _alUDdUro_ /	/ Dk /	adquirir	/ _aDkirIR_ /
/ Dg /	alud ganado	/ _alUDganADo_ /	/ DB /	adverbio	/ _aDBErBjo_ /
/ Dm /	administrar	/ _aDministrAR_ /	/ Dn /	alud negro	/ _alUDnEGro_ /
/ DJ /	alud ñoño	/ _alUDJOJo_ /	/ Df /	alud feo	/ _alUDfEo_ /
/ DT /	alud zapato	/ _alUDTapAto_ /	/ Ds /	adsorción	/ _aDsorTjOn_ /
/ DC /	ardid chulo	/ _arDIDCUlo_ /	/ Dx /	David jota	/ _daBIDxOta_ /
/ DL /	alud llave	/ _alUDLABe_ /	/ DR /	alud raro	/ _alUDRAro_ /
/ Dj /	adyacente	/ _aDjaTEnte_ /	/ DI /	medir	/ _meDIR_ /
/ DE /	madera	/ _maDEra_ /	/ Da /	moda	/ _mODa_ /
/ Do /	dado	/ _dADo_ /	/ Du /	caducar	/ _kaDukAR_ /
/ D_ /	David	/ _daBID_ /	/ sp /	espejo	/ _espExo_ /
/ st /	lástima	/ _lAstima_ /	/ sk /	descaro	/ _deskAro_ /
/ sm /	esmero	/ _esmEro_ /	/ sn /	desnivel	/ _desniBEl_ /
/ sJ /	es ñoño	/ _EsJOJo_ /	/ sB /	esbirro	/ _esBIRo_ /
/ sf /	esfera	/ _esfEra_ /	/ sT /	escena	/ _esTEna_ /
/ sD /	esdrújulo	/ _esDrUxulo_ /	/ ss /	es suelo	/ _EsswElo_ /
/ sC /	es chato	/ _EsCAto_ /	/ sx /	es juzgado	/ _EsxuTGADo_ /
/ sG /	desguace	/ _desGwATe_ /	/ sl /	su isla	/ _sUIsla_ /
/ sL /	esas llaves	/ _EsasLABes_ /	/ sR /	desratización	/ _desRatiTaTjOn_ /
/ sI /	sitio	/ _sItjo_ /	/ se /	sereno	/ _serEno_ /
/ sa /	salida	/ _salIDa_ /	/ sO /	sobre	/ _sOBre_ /
/ sU /	susto	/ _sUsto_ /	/ s_ /	eses	/ _Eses_ /
/ CI /	chiste	/ _CIste_ /	/ Ce /	leche	/ _lECe_ /
/ Ca /	hacha	/ _ACa_ /	/ Co /	cacho	/ _kACo_ /
/ CU /	chucho	/ _CUCo_ /	/ xp /	reloj pared	/ _RelOxparED_ /
/ xb /	reloj bueno	/ _RelOxbwEno_ /	/ xt /	reloj tapado	/ _RelOxtapADo_ /
/ xd /	reloj duro	/ _RelOxdUro_ /	/ xk /	reloj caro	/ _RelOxkAro_ /
/ xg /	reloj ganado	/ _RelOxganADo_ /	/ xm /	reloj malo	/ _RelOxmAlo_ /
/ xn /	reloj nuevo	/ _RelOxnwEBo_ /	/ xJ /	reloj ñoño	/ _RelOxJOJo_ /
/ xf /	reloj feo	/ _RelOxfEo_ /	/ xT /	reloj cerca	/ _RelOxTErka_ /
/ xs /	reloj suena	/ _RelOxswEna_ /	/ xC /	reloj chato	/ _RelOxCAto_ /
/ xx /	reloj joven	/ _RelOxxOBen_ /	/ xl /	reloj listo	/ _RelOxlIsto_ /
/ xL /	reloj llavero	/ _RelOxLaBEro_ /	/ xR /	reloj ruidoso	/ _RelOxRwiDOso_ /
/ xj /	reloj hierro	/ _RelOxjERo_ /	/ xE /	jefe	/ _xEfe_ /
/ xA /	jarra	/ _xARa_ /	/ xO /	jota	/ _xOta_ /
/ xu /	juntar	/ _xuntAR_ /	/ x_ /	reloj	/ _RelOx_ /
/ kB /	macba	/ _mAkBa_ /	/ kt /	actuar	/ _aktwAR_ /
/ Gd /	magdalena	/ _maGdalEna_ /	/ km /	acme	/ _Akme_ /
/ Gn /	magnético	/ _maGnEtiko_ /	/ kT /	acción	/ _akTjOn_ /
/ ks /	óxido	/ _OksiDo_ /	/ Gi /	águila	/ _AGila_ /
/ GE /	la guerra	/ _lAGERa_ /	/ GA /	la gala	/ _lAGAla_ /
/ Go /	algo	/ _AlGo_ /	/ GU /	alguno	/ _alGUno_ /
/ lp /	col pisar	/ _kOlpisAR_ /	/ lt /	alto	/ _Alto_ /
/ lD /	aldea	/ _alDEa_ /	/ lk /	alcance	/ _alkAnTe_ /
/ lm /	alma	/ _Alma_ /	/ ln /	malnacido	/ _malnaTIDo_ /
/ lJ /	sal ñoño	/ _sAlJOJo_ /	/ lB /	alba	/ _AlBa_ /
/ lf /	alfiler	/ _alfilER_ /	/ lT /	alce	/ _AlTe_ /
/ ls /	el suelo	/ _ElswElo_ /	/ lC /	Elche	/ _ElCe_ /
/ lx /	el joven	/ _ElxOBen_ /	/ lG /	alguno	/ _alGUno_ /
/ ll /	al lado	/ _AllADo_ /	/ lL /	sal llena	/ _sAlLEna_ /
/ lR /	alrededor	/ _alReDeDOR_ /	/ lI /	libro	/ _lIBro_ /
/ lE /	lema	/ _lEma_ /	/ lA /	lástima	/ _lAstima_ /
/ lo /	colocar	/ _kolokAR_ /	/ lU /	luna	/ _lUna_ /
/ l_ /	col	/ _kOl_ /	/ LI /	allí	/ _aLI_ /
/ Le /	calle	/ _kALe_ /	/ LA /	llave	/ _LABe_ /
/ Lo /	llorar	/ _LorAR_ /	/ LU /	lluvia	/ _LUBja_ /
/ RI /	risa	/ _RIsa_ /	/ Re /	reloj	/ _RelOx_ /
/ RA /	rata	/ _RAta_ /	/ RO /	roca	/ _ROka_ /
/ RU /	ruso	/ _RUso_ /	/ R_ /	tocar	/ _tokAR_ /

Tabla B.5: Lista de difonemas y trifonemas (III).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ rp /	arpón	/ _arpOn_ /	/ rt /	arte	/ _Arte_ /
/ rk /	arca	/ _Arka_ /	/ rm /	armario	/ _armArjo_ /
/ rn /	barniz	/ _barnIT_ /	/ rJ /	lugar ñoño	/ _luGArJOJo_ /
/ rB /	árbol	/ _ArBol_ /	/ rf /	garfio	/ _gArfjo_ /
/ rT /	arce	/ _ArTe_ /	/ rD /	arder	/ _arDER_ /
/ rs /	arsénico	/ _arsEniko_ /	/ rC /	archivar	/ _arCiBAR_ /
/ rx /	forjar	/ _forxAR_ /	/ rG /	argón	/ _arGOn_ /
/ rl /	arlequín	/ _arlekIn_ /	/ rL /	lugar lleno	/ _luGArLEno_ /
/ rR /	llevar rato	/ _LeBArRAto_ /	/ ri /	árido	/ _AriDo_ /
/ rE /	arena	/ _arEna_ /	/ ra /	tocara	/ _tokAra_ /
/ ro /	oro	/ _Oro_ /	/ rU /	oruga	/ _orUGa_ /
/ jp /	hoy puedo	/ _OjpwEDo_ /	/ jt /	hay te	/ _AjtE_ /
/ jk /	hay café	/ _AjkafE_ /	/ jm /	hay mano	/ _AjmAno_ /
/ jn /	hay niño	/ _AjnIJo_ /	/ jJ /	hoy ñoño	/ _OjJOJo_ /
/ jB /	hoy buenas	/ _OjBwEnas_ /	/ jf /	hoy fieras	/ _OjfjEras_ /
/ jT /	hoy cielo	/ _OjTjElo_ /	/ jD /	hoy duerme	/ _OjDwErme_ /
/ js /	hoy sale	/ _OjsAle_ /	/ jC /	hay chistes	/ _AjCIstes_ /
/ jx /	hay gente	/ _AjxEnte_ /	/ jG /	hay guerra	/ _AjGERa_ /
/ jl /	hay lío	/ _AjlIo_ /	/ jL /	hay llaves	/ _AjLABes_ /
/ jR /	hay rata	/ _AjRAta_ /	/ jr /	aire	/ _Ajre_ /
/ jj /	hoy hierve	/ _OjjErBe_ /	/ jw /	hoyuelo	/ _ojwElo_ /
/ jI /	hay isla	/ _AjIsla_ /	/ jE /	hiena	/ _jEna_ /
/ jA /	hiato	/ _jAto_ /	/ jO /	iota	/ _jOta_ /
/ jU /	Iuta	/ _jUta_ /	/ j_ /	hoy	/ _Oj_ /
/ wp /	aupar	/ _awpAR_ /	/ wt /	auto	/ _Awto_ /
/ wk /	eucalipto	/ _ewkalIpto_ /	/ wm /	aumento	/ _awmEnto_ /
/ wn /	eunuco	/ _ewnUko_ /	/ wN /	aunque	/ _AwNke_ /
/ wB /	aubernés	/ _awBernEs_ /	/ wf /	euforia	/ _ewfOrja_ /
/ wT /	leucemia	/ _lewTEmja_ /	/ wD /	laude	/ _lAwDe_ /
/ ws /	ausencia	/ _awsEnTja_ /	/ eu /	suele huchear	/ _swEleuCeAR_ /
/ wx /	auge	/ _Awxe_ /	/ wG /	augurar	/ _awGurAR_ /
/ wl /	aula	/ _Awla_ /	/ wL /	aullar	/ _awLAR_ /
/ wr /	aura	/ _Awra_ /	/ wI /	ruin	/ _RwIn_ /
/ wE /	duelo	/ _dwElo_ /	/ wA /	dual	/ _dwAl_ /
/ wo /	duodécimo	/ _dwoDETimo_ /	/ Ip /	hipo	/ _Ipo_ /
/ It /	sitio	/ _sItjo_ /	/ Ik /	dique	/ _dIke_ /
/ im /	imagen	/ _imAxen_ /	/ in /	incendio	/ _inTEnDjo_ /
/ IJ /	niño	/ _nIJo_ /	/ iN /	incalculable	/ _iNkalkulABle_ /
/ iM /	infierno	/ _iMfjErno_ /	/ IB /	alivio	/ _alIBjo_ /
/ if /	calificable	/ _kalifikABle_ /	/ iT /	izar	/ _iTAR_ /
/ ID /	ídolo	/ _IDolo_ /	/ Is /	isla	/ _Isla_ /
/ Ix /	hijo	/ _Ixo_ /	/ IG /	higo	/ _IGo_ /
/ Il /	mil	/ _mIl_ /	/ IL /	silla	/ _sILa_ /
/ iR /	irradiar	/ _iRaDjAR_ /	/ Ir /	ira	/ _Ira_ /
/ Ij /	salí hiato	/ _salIjAto_ /	/ Iw /	rompí hueso	/ _RompIwEso_ /
/ II /	salí isla	/ _salIIsla_ /	/ Ie /	comí encima	/ _komIenTIma_ /
/ Ia /	comí avión	/ _komIaBjOn_ /	/ IO /	comí oso	/ _komIOso_ /
/ IU /	comí uva	/ _komIUBa_ /	/ I_ /	salí	/ _salI_ /
/ ep /	epílogo	/ _epIloGo_ /	/ et /	eterno	/ _etErno_ /
/ Ek /	eco	/ _Eko_ /	/ em /	emigrar	/ _emiGrAR_ /
/ en /	encima	/ _enTIma_ /	/ EJ /	eñe	/ _EJe_ /
/ eN /	encuentro	/ _eNkwEntro_ /	/ EM /	énfasis	/ _EMfasis_ /
/ eB /	evitar	/ _eBitAR_ /	/ ef /	efecto	/ _efEkto_ /
/ ET /	cerezo	/ _TerETo_ /	/ ED /	dedo	/ _dEDo_ /
/ es /	estado	/ _estADo_ /	/ eC /	hechizo	/ _eCITo_ /
/ ex /	ejemplo	/ _exEmplo_ /	/ eG /	negar	/ _neGAR_ /
/ el /	elefante	/ _elefAnte_ /	/ EL /	bella	/ _bELa_ /
/ ER /	cerro	/ _TERo_ /	/ Er /	cera	/ _TEra_ /
/ Ej /	ley	/ _lEj_ /	/ Ew /	reuma	/ _REwma_ /
/ eI /	sale isla	/ _sAleIsla_ /	/ ee /	calle estrecha	/ _kALeestrECa_ /
/ eA /	calle ancha	/ _kALeAnCa_ /	/ eo /	calle oscura	/ _kALeoskUra_ /
/ eU /	calle húmeda	/ _kALeUmeDa_ /	/ e_ /	calle	/ _kALe_ /
/ ap /	capitán	/ _kapitAn_ /	/ At /	cata	/ _kAta_ /
/ ak /	aquí	/ _akI_ /	/ Am /	fama	/ _fAma_ /

Tabla B.6: Lista de difonemas y trifonemas (IV).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ An /	Ana	/ _Ana_ /	/ AJ /	año	/ _AJo_ /
/ AN /	ángulo	/ _ANGulo_ /	/ aM /	anfibio	/ _aMfIBjo_ /
/ aB /	abierto	/ _aBjErto_ /	/ af /	afeitar	/ _afejtAR_ /
/ aT /	aceite	/ _aTEjte_ /	/ AD /	hada	/ _ADa_ /
/ As /	asta	/ _Asta_ /	/ AC /	hacha	/ _ACa_ /
/ Ax /	ajo	/ _Axo_ /	/ AG /	ágora	/ _AGora_ /
/ Al /	alma	/ _Alma_ /	/ AL /	calle	/ _kALe_ /
/ aR /	arrojar	/ _aRoxAR_ /	/ Ar /	cara	/ _kAra_ /
/ Aj /	aire	/ _Ajre_ /	/ Aw /	aura	/ _Awra_ /
/ aI /	esa isla	/ _EsaIsla_ /	/ aE /	aéreo	/ _aEreo_ /
/ aA /	esa asa	/ _EsaAsa_ /	/ ao /	ahogar	/ _aoGAR_ /
/ aU /	esa única	/ _EsaUnika_ /	/ a_ /	cesta	/ _TEsta_ /
/ Op /	ópera	/ _Opera_ /	/ Ot /	otro	/ _Otro_ /
/ ok /	ocaso	/ _okAso_ /	/ Om /	hombro	/ _OmBro_ /
/ On /	once	/ _OnTe_ /	/ OJ /	ñoño	/ _JOJo_ /
/ ON /	con que	/ _kONkE_ /	/ oM /	conferencia	/ _koMferEnTja_ /
/ oB /	oboe	/ _oBOe_ /	/ of /	ofrecer	/ _ofreTER_ /
/ oT /	tocino	/ _toTIno_ /	/ OD /	odio	/ _ODjo_ /
/ Os /	ostra	/ _Ostra_ /	/ OC /	ocho	/ _OCo_ /
/ Ox /	ojo	/ _Oxo_ /	/ oG /	hogar	/ _oGAR_ /
/ ol /	colocar	/ _kolokAR_ /	/ OL /	olla	/ _OLa_ /
/ OR /	corro	/ _kORo_ /	/ Or /	oro	/ _Oro_ /
/ Oj /	boina	/ _bOjna_ /	/ ow /	ojo hueco	/ _OxowEko_ /
/ oI /	ojo isla	/ _OxoIsla_ /	/ Oe /	oboe	/ _oBOe_ /
/ oA /	coala	/ _koAla_ /	/ oO /	cojo ocho	/ _kOxoOCo_ /
/ oU /	cojo uva	/ _kOxoUBa_ /	/ o_ /	ducho	/ _dUCo_ /
/ Up /	grupo	/ _grUpo_ /	/ Ut /	bruto	/ _brUto_ /
/ Uk /	nuca	/ _nUka_ /	/ Um /	suma	/ _sUma_ /
/ Un /	uno	/ _Uno_ /	/ UJ /	uña	/ _UJa_ /
/ UN /	nunca	/ _nUNka_ /	/ UM /	un faro	/ _UMfAro_ /
/ UB /	tubo	/ _tUBo_ /	/ uf /	bufón	/ _bufOn_ /
/ UT /	buzo	/ _bUTo_ /	/ UD /	duda	/ _dUDa_ /
/ us /	usar	/ _usAR_ /	/ UC /	hucha	/ _UCa_ /
/ ux /	agujero	/ _aGuxEro_ /	/ uG /	lugar	/ _luGAR_ /
/ ul /	enjaular	/ _eNxawlAR_ /	/ UL /	su llave	/ _sULABe_ /
/ UR /	hurra	/ _URa_ /	/ Ur /	cura	/ _kUra_ /
/ Uj /	su hiena	/ _sUjEna_ /	/ Uw /	su huevo	/ _sUwEBo_ /
/ UI /	su isla	/ _sUIsla_ /	/ Ue /	su edad	/ _sUeDAD_ /
/ UA /	su arca	/ _sUArka_ /	/ UO /	su oso	/ _sUOso_ /
/ UU /	su única	/ _sUUnika_ /	/ U_ /	su	/ _sU_ /
/ _p /	pisar	/ _pisAR_ /	/ _b /	bien	/ _bjEn_ /
/ _t /	todo	/ _tODo_ /	/ _d /	diente	/ _djEnte_ /
/ _k /	casa	/ _kAsa_ /	/ _g /	guante	/ _gwAnte_ /
/ _m /	mesa	/ _mEsa_ /	/ _n /	nadie	/ _nADje_ /
/ _J /	ñoño	/ _JOJo_ /	/ _f /	firma	/ _fIrma_ /
/ _T /	zapato	/ _TapAto_ /	/ _s /	suelo	/ _swElo_ /
/ _C /	chapa	/ _CApa_ /	/ _x /	jarrón	/ _xaROn_ /
/ _l /	lema	/ _lEma_ /	/ _L /	llamar	/ _LamAR_ /
/ _R /	risa	/ _RIsa_ /	/ _j /	hiena	/ _jEna_ /
/ _w /	hueco	/ _wEko_ /	/ _I /	ira	/ _Ira_ /
/ _e /	encima	/ _enTIma_ /	/ _a /	acabar	/ _akaBAR_ /
/ _o /	ortiga	/ _ortIGa_ /	/ _u /	humano	/ _umAno_ /
/ plj /	pliegue	/ _pljEGe_ /	/ plI /	coplilla	/ _koplILa_ /
/ plE /	pleno	/ _plEno_ /	/ plA /	Calaf playa	/ _kalAfplAja_ /
/ plO /	plomo	/ _plOmo_ /	/ plU /	pluma	/ _plUma_ /
/ prj /	prioridad	/ _prjoriDAD_ /	/ prw /	prueba	/ _prwEBa_ /
/ prI /	prisa	/ _prIsa_ /	/ prE /	previo	/ _prEBjo_ /
/ prA /	Praga	/ _prAGa_ /	/ pro /	provocar	/ _proBokAR_ /
/ pru /	prusiano	/ _prusjAno_ /	/ pjE /	pie	/ _pjE_ /
/ pjA /	piano	/ _pjAno_ /	/ pjO /	piojo	/ _pjOxo_ /
/ pwE /	puerta	/ _pwErta_ /	/ pwA /	puar	/ _pwAR_ /
/ blI /	blinco	/ _blINko_ /	/ blE /	blécua	/ _blEkwa_ /
/ blA /	blanco	/ _blANko_ /	/ blO /	bloque	/ _blOke_ /
/ blU /	blusa	/ _blUsa_ /	/ brj /	brioso	/ _brjOso_ /

Tabla B.7: Lista de difonemas y trifonemas (V).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ brI /	brisa	/ _brIsa_ /	/ bre /	bretón	/ _bretOn_ /
/ brA /	brazo	/ _brATo_ /	/ brO /	bronca	/ _brONka_ /
/ brU /	bruja	/ _brUxa_ /	/ bjE /	viejo	/ _bjExo_ /
/ bjA /	vial	/ _bjAl_ /	/ bjO /	biónico	/ _bjOniko_ /
/ bwI /	buitre	/ _bwItre_ /	/ bwE /	bueno	/ _bwEno_ /
/ bwA /	buana	/ _bwAna_ /	/ bwo /	buhonero	/ _bwonEro_ /
/ trj /	trial	/ _trjAl_ /	/ trw /	truhán	/ _trwAn_ /
/ tri /	trifonema	/ _trifonEma_ /	/ trE /	trenza	/ _trEnTa_ /
/ trA /	tráquea	/ _trAkea_ /	/ tro /	trotar	/ _trotAR_ /
/ trU /	trufa	/ _trUfa_ /	/ tjE /	tiene	/ _tjEne_ /
/ tjA /	Tiana	/ _tjAna_ /	/ tjo /	patio	/ _pAtjo_ /
/ twI /	twist	/ _twIst_ /	/ twE /	tuétano	/ _twEtano_ /
/ twa /	perpetua	/ _perpEtwa_ /	/ two /	fatuo	/ _fAtwo_ /
/ drw /	druida	/ _drwIDa_ /	/ dri /	driblar	/ _driBlAR_ /
/ dre /	drenar	/ _drenAR_ /	/ drA /	drama	/ _drAma_ /
/ dro /	dromedario	/ _dromeDArjo_ /	/ drU /	drum	/ _drUm_ /
/ djE /	diedro	/ _djEDro_ /	/ djA /	diáfano	/ _djAfano_ /
/ djO /	Dios	/ _djOs_ /	/ dwE /	duerme	/ _dwErme_ /
/ dwa /	dualidad	/ _dwaliDAD_ /	/ dwo /	duodécimo	/ _dwoDETimo_ /
/ klj /	cliente	/ _kljEnte_ /	/ klw /	clueca	/ _klwEka_ /
/ klI /	clima	/ _klIma_ /	/ kle /	cleptómano	/ _kleptOmano_ /
/ klA /	clara	/ _klAra_ /	/ klO /	cloro	/ _klOro_ /
/ klU /	club	/ _klUB_ /	/ krj /	crianza	/ _krjAnTa_ /
/ krw /	cruenta	/ _krwEnta_ /	/ kri /	cristal	/ _kristAl_ /
/ krE /	cresta	/ _krEsta_ /	/ krA /	cráneo	/ _krAneo_ /
/ krO /	cromo	/ _krOmo_ /	/ krU /	cruz	/ _krUT_ /
/ kjE /	quieto	/ _kjEto_ /	/ kja /	acequia	/ _aTEkja_ /
/ kjo /	obsequio	/ _oBsEkjo_ /	/ kwi /	cuidar	/ _kwiDAR_ /
/ kwE /	cuesta	/ _kwEsta_ /	/ kwa /	inocua	/ _inOkwa_ /
/ kwo /	inocuo	/ _inOkwo_ /	/ gli /	glicerina	/ _gliTerIna_ /
/ glE /	gleba	/ _glEBa_ /	/ gla /	glaucoma	/ _glawkOma_ /
/ glO /	globo	/ _glOBo_ /	/ glU /	glúteo	/ _glUteo_ /
/ grj /	grieta	/ _grjEta_ /	/ grw /	grueso	/ _grwEso_ /
/ grI /	grima	/ _grIma_ /	/ gre /	gregario	/ _greGArjo_ /
/ grA /	grasa	/ _grAsa_ /	/ gro /	grotesco	/ _grotEsko_ /
/ grU /	gruta	/ _grUta_ /	/ gjA /	guiarse	/ _gjArse_ /
/ gjO /	guión	/ _gjOn_ /	/ gwE /	güelfo	/ _gwElfo_ /
/ gwA /	guante	/ _gwAnte_ /	/ mjE /	miércoles	/ _mjErkoles_ /
/ mja /	academia	/ _akaDEmja_ /	/ mjo /	miopía	/ _mjopIa_ /
/ mwE /	mueca	/ _mwEka_ /	/ mwa /	Ermua	/ _Ermwa_ /
/ njE /	nieve	/ _njEBe_ /	/ njA /	Niágara	/ _njAGara_ /
/ njO /	reunión	/ _RewnjOn_ /	/ nwE /	nuera	/ _nwEra_ /
/ nwA /	anual	/ _anwAl_ /	/ nwO /	sinuoso	/ _sinwOso_ /
/ Blj /	Biblia	/ _bIBlja_ /	/ Bli /	bíblico	/ _bIBliko_ /
/ Ble /	sable	/ _sABle_ /	/ BlA /	hablar	/ _aBlAR_ /
/ BlO /	doblón	/ _doBlOn_ /	/ Blu /	ablución	/ _aBluTjOn_ /
/ Brj /	abriendo	/ _aBrjEnDo_ /	/ BrI /	abrir	/ _aBrIR_ /
/ Bre /	abreviar	/ _aBreBjAR_ /	/ Bra /	culebra	/ _kulEBra_ /
/ Bro /	abrochar	/ _aBroCAR_ /	/ Bru /	abrumar	/ _aBrumAR_ /
/ BjE /	abierto	/ _aBjErto_ /	/ Bja /	aviación	/ _aBjaTjOn_ /
/ BjO /	avión	/ _aBjOn_ /	/ BwE /	abuelo	/ _aBwElo_ /
/ flw /	superfluo	/ _supErflwo_ /	/ fli /	afligir	/ _aflixIR_ /
/ flE /	flecha	/ _flECa_ /	/ flA /	flaco	/ _flAko_ /
/ flO /	flor	/ _flOR_ /	/ flU /	fluca	/ _flUka_ /
/ frj /	friolera	/ _frjolEra_ /	/ frw /	fruición	/ _frwiTjOn_ /
/ frI /	friso	/ _frIso_ /	/ frE /	fresa	/ _frEsa_ /
/ frA /	frase	/ _frAse_ /	/ fro /	afro	/ _Afro_ /
/ frU /	fruta	/ _frUta_ /	/ fjE /	fiera	/ _fjEra_ /
/ fjA /	fiar	/ _fjAR_ /	/ fjo /	zafio	/ _TAfjo_ /
/ fwI /	fuimos	/ _fwImos_ /	/ fwE /	fuero	/ _fwEro_ /
/ fwa /	fuagrás	/ _fwaGrAs_ /	/ TjE /	cielo	/ _TjElo_ /
/ TjA /	comercial	/ _komerTjAl_ /	/ TjO /	acción	/ _akTjOn_ /
/ TwE /	zueco	/ _TwEko_ /	/ Drw /	la druida	/ _lADrwIDa_ /
/ DrI /	padrino	/ _paDrIno_ /	/ Dre /	pudre	/ _pUDre_ /

Tabla B.8: Lista de difonemas y trifonemas (VI).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ Dra /	hiedra	/ _jEDra_ /	/ Dro /	diedro	/ _djEDro_ /
/ DrU /	esdrújula	/ _esDrUxula_ /	/ Dje /	nadie	/ _nADje_ /
/ Dja /	comedia	/ _komEDja_ /	/ DjO /	adiós	/ _aDjOs_ /
/ DwI /	abduir	/ _aBdwIR_ /	/ Dwa /	ardua	/ _ArDwa_ /
/ Dwe /	adueñar	/ _aDweJAR_ /	/ Dwo /	arduo	/ _ArDwo_ /
/ sjE /	siesta	/ _sjEsta_ /	/ sjA /	Siam	/ _sjAm_ /
/ sjo /	sionista	/ _sjonIsta_ /	/ swI /	suizo	/ _swITo_ /
/ swE /	suelo	/ _swElo_ /	/ swA /	suave	/ _swABe_ /
/ xjE /	ujier	/ _uxjER_ /	/ xjA /	colegial	/ _kolexjAl_ /
/ xjo /	colegio	/ _kolExjo_ /	/ xwI /	juicio	/ _xwITjo_ /
/ xwE /	jueves	/ _xwEBes_ /	/ xwA /	ajuar	/ _axwAR_ /
/ GlE /	iglesia	/ _iGlEsja_ /	/ Gla /	sigla	/ _sIGla_ /
/ Glo /	siglo	/ _sIGlo_ /	/ GlU /	iglú	/ _iGlU_ /
/ Grj /	agria	/ _AGrja_ /	/ Grw /	incongruencia	/ _iNkoNGrwEnTja_ /
/ Gri /	agricultura	/ _aGrikultUra_ /	/ GrE /	agreste	/ _aGrEste_ /
/ GrA /	agravio	/ _aGrABjo_ /	/ GrO /	agrónomo	/ _aGrOnomo_ /
/ Gru /	agrupar	/ _aGrupAR_ /	/ Gje /	alguien	/ _AlGjen_ /
/ GjA /	se guiado	/ _sEGjADo_ /	/ GjO /	se guión	/ _sEGjOn_ /
/ Gwe /	antigüedad	/ _antiGweDAD_ /	/ Gwa /	antigua	/ _antIGwa_ /
/ Gwo /	antiguo	/ _antIGwo_ /	/ lje /	alienar	/ _aljenAR_ /
/ ljA /	liana	/ _ljAna_ /	/ ljO /	lioso	/ _ljOso_ /
/ lwE /	luengo	/ _lwENGo_ /	/ lwA /	baluarte	/ _balwArte_ /
/ lwO /	valuoso	/ _balwOso_ /	/ lwI /	Luís	/ _lwIs_ /
/ LwE /	llueve	/ _LwEBe_ /	/ RjE /	riesgo	/ _RjEsGo_ /
/ RjA /	riada	/ _RjADa_ /	/ RjO /	carrión	/ _kaRjOn_ /
/ RwI /	ruido	/ _RwIDo_ /	/ RwE /	ruego	/ _RwEGo_ /
/ RwA /	Ruanda	/ _RwAnDa_ /	/ rjE /	ariete	/ _arjEte_ /
/ rja /	diaria	/ _djArja_ /	/ rjo /	diario	/ _djArjo_ /
/ CjA /	Chiapas	/ _CjApas_ /	/ Cwa /	Quechua	/ _kECwa_ /
/ iC /	fichar	/ _fiCAR_ /	/ IC /	dicho	/ _dICo_ /
/ pt /	apto	/ _Apto_ /	/ pI /	pista	/ _pIsta_ /
/ pe /	peseta	/ _pesEta_ /	/ pa /	capa	/ _kApa_ /
/ po /	podar	/ _poDAR_ /	/ pu /	pulir	/ _pulIR_ /
/ bi /	biberón	/ _biBerOn_ /	/ bE /	bebe	/ _bEBe_ /
/ bA /	barco	/ _bArko_ /	/ bo /	bobada	/ _boBADa_ /
/ bU /	buzo	/ _bUTo_ /	/ tI /	timo	/ _tImo_ /
/ tE /	té	/ _tE_ /	/ ta /	lata	/ _lAta_ /
/ tO /	todo	/ _tODo_ /	/ tu /	aturdir	/ _aturDIR_ /
/ dI /	dicho	/ _dICo_ /	/ de /	dedal	/ _deDAl_ /
/ da /	dañar	/ _daJAR_ /	/ dO /	dos	/ _dOs_ /
/ du /	dudar	/ _duDAR_ /	/ ki /	quilate	/ _kilAte_ /
/ ke /	saque	/ _sAke_ /	/ ka /	calar	/ _kalAR_ /
/ ko /	coger	/ _koxER_ /	/ ku /	acusar	/ _akusAR_ /
/ gI /	guiso	/ _gIso_ /	/ ge /	guerrero	/ _geREro_ /
/ ga /	gastar	/ _gastAR_ /	/ go /	golear	/ _goleAR_ /
/ gu /	gustar	/ _gustAR_ /	/ mI /	camisa	/ _kamIsa_ /
/ me /	medir	/ _meDIR_ /	/ ma /	cama	/ _kAma_ /
/ mo /	cómo	/ _kOmo_ /	/ mu /	mujer	/ _muxER_ /
/ ni /	anidar	/ _aniDAR_ /	/ nE /	negro	/ _nEGro_ /
/ nA /	nada	/ _nADa_ /	/ nO /	nómada	/ _nOmaDa_ /
/ Ji /	cañizal	/ _kaJiTAl_ /	/ Je /	cañería	/ _kaJerIa_ /
/ JA /	cañada	/ _kaJADa_ /	/ JO /	cañón	/ _kaJOn_ /
/ Ju /	cañu	/ _kAJu_ /	/ Bi /	avisar	/ _aBisAR_ /
/ Be /	abejorro	/ _aBexORo_ /	/ BA /	reválida	/ _ReBAliDa_ /
/ BO /	arbóreo	/ _arBOreo_ /	/ BU /	abuso	/ _aBUso_ /
/ fi /	firmar	/ _firmAR_ /	/ fe /	feliz	/ _felIT_ /
/ fa /	falaz	/ _falAT_ /	/ fo /	formar	/ _formAR_ /
/ fu /	fundir	/ _funDIR_ /	/ Ti /	cineasta	/ _TineAsta_ /
/ Te /	encestar	/ _enTestAR_ /	/ TA /	cazar	/ _kaTAR_ /
/ TO /	cazó	/ _kaTO_ /	/ Tu /	zurrón	/ _TuROn_ /
/ Di /	adivino	/ _aDiBIno_ /	/ De /	aderezo	/ _aDerETo_ /
/ DA /	nadar	/ _naDAR_ /	/ DO /	ardor	/ _arDOR_ /
/ DU /	adúltero	/ _aDUltero_ /	/ si /	sitiar	/ _sitjAR_ /
/ sE /	ser	/ _sER_ /	/ sA /	sabio	/ _sABjo_ /

Tabla B.9: Lista de difonemas y trifonemas (VII).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ so /	sobrar	/ _soBrAR_ /	/ su /	sultán	/ _sultAn_ /
/ xI /	cojín	/ _koxIn_ /	/ xe /	jerez	/ _xerET_ /
/ xa /	jarrón	/ _xaROn_ /	/ xo /	jornada	/ _xornADa_ /
/ xU /	junta	/ _xUnta_ /	/ GI /	seguir	/ _seGIR_ /
/ Ge /	aguerrido	/ _aGeRIDo_ /	/ Ga /	ágape	/ _AGape_ /
/ GO /	agobio	/ _aGOBjo_ /	/ Gu /	regular	/ _ReGulAR_ /
/ li /	librar	/ _liBrAR_ /	/ le /	leñador	/ _leJaDOR_ /
/ la /	lastimar	/ _lastimAR_ /	/ lO /	loco	/ _lOko_ /
/ lu /	lunar	/ _lunAR_ /	/ Li /	gallináceo	/ _gaLinATeo_ /
/ LE /	callé	/ _kaLE_ /	/ La /	llavero	/ _LaBEro_ /
/ LO /	llora	/ _LOra_ /	/ Lu /	lluvioso	/ _LuBjOso_ /
/ Ri /	arrimar	/ _aRimAR_ /	/ RE /	arresto	/ _aREsto_ /
/ Ra /	radar	/ _RaDAR_ /	/ Ro /	rogar	/ _RoGAR_ /
/ Ru /	arrugar	/ _aRuGAR_ /	/ rI /	arista	/ _arIsta_ /
/ re /	arenal	/ _arenAl_ /	/ rA /	arácnido	/ _arAkniDo_ /
/ rO /	aroma	/ _arOma_ /	/ ru /	cirujano	/ _TiruxAno_ /
/ ji /	hay isleño	/ _AjislEJo_ /	/ je /	cambien	/ _kAmBjen_ /
/ ja /	recia	/ _RETja_ /	/ jo /	necio	/ _nETjo_ /
/ ju /	ciudad	/ _TjuDAD_ /	/ wi /	ruinoso	/ _RwinOso_ /
/ we /	pueblerino	/ _pweBlerIno_ /	/ wa /	blécua	/ _blEkwa_ /
/ wO /	arduoso	/ _arDwOso_ /	/ ip /	hipoglucemia	/ _ipoGluTEmja_ /
/ it /	sitiado	/ _sitjADo_ /	/ ik /	picar	/ _pikAR_ /
/ Im /	tímido	/ _tImiDo_ /	/ In /	cínico	/ _TIniko_ /
/ iJ /	piñón	/ _piJOn_ /	/ IN /	cinco	/ _TINko_ /
/ IM /	ínfimo	/ _IMfimo_ /	/ IB /	víbora	/ _bIBora_ /
/ iD /	ideal	/ _iDeAl_ /	/ is /	pisar	/ _pisAR_ /
/ ix /	fijar	/ _fixAR_ /	/ iG /	higuera	/ _iGEra_ /
/ il /	militar	/ _militAR_ /	/ iL /	ensillar	/ _ensiLAR_ /
/ IR /	mirra	/ _mIRa_ /	/ ir /	mirar	/ _mirAR_ /
/ ij /	casi yogur	/ _kAsijoGUR_ /	/ Ii /	comí inspirado	/ _komIinspirADo_ /
/ ie /	casi está	/ _kAsiestA_ /	/ iE /	casi ella	/ _kAsiELa_ /
/ IE /	comí esto	/ _komIEsto_ /	/ iA /	casi años	/ _kAsiAJos_ /
/ ia /	casi allá	/ _kAsiaLA_ /	/ IA /	comí hasta	/ _komIAsta_ /
/ IO /	comí ocho	/ _komIOCo_ /	/ iO /	casi oro	/ _kAsiOro_ /
/ Io /	comí osobuco	/ _komIosoBUko_ /	/ iU /	Mali único	/ _mAliUniko_ /
/ Iu /	mi universidad	/ _mIuniBersiDAD_ /	/ iu /	Mali unido	/ _mAliunIDo_ /
/ i_ /	Mali	/ _mAli_ /	/ Ep /	cepa	/ _TEpa_ /
/ Et /	cetro	/ _TEtro_ /	/ ek /	detector	/ _detektOR_ /
/ Em /	tema	/ _tEma_ /	/ En /	cena	/ _TEna_ /
/ eJ /	leñador	/ _leJaDOR_ /	/ eN /	encuentro	/ _eNkwEntro_ /
/ eM /	enfadar	/ _eMfaDAR_ /	/ EB /	ébano	/ _EBano_ /
/ Ef /	encéfalo	/ _enTEfalo_ /	/ ET /	heces	/ _ETes_ /
/ eD /	dedal	/ _deDAl_ /	/ Es /	este	/ _Este_ /
/ EC /	pecho	/ _pECo_ /	/ Ex /	teja	/ _tExa_ /
/ EG /	pega	/ _pEGa_ /	/ El /	tela	/ _tEla_ /
/ eL /	belleza	/ _beLETa_ /	/ eR /	cerrar	/ _TeRAR_ /
/ er /	encerar	/ _enTerAR_ /	/ ej /	come hierro	/ _kOmejERo_ /
/ ew /	feudal	/ _fewDAl_ /	/ EI /	tomé isla	/ _tomEIsla_ /
/ ei /	dale ideas	/ _dAleiDEas_ /	/ Ei /	té inglés	/ _tEiNGlEs_ /
/ EE /	tomé esto	/ _tomEEsto_ /	/ Ee /	tomé estrella	/ _tomEestrELa_ /
/ eE /	calle esta	/ _kALeEsta_ /	/ ea /	calle alargada	/ _kALealarGADa_ /
/ EA /	tomé algo	/ _tomEAlGo_ /	/ Ea /	tomé alguna	/ _tomEalGUna_ /
/ eO /	calle ocho	/ _kALeOCo_ /	/ EO /	tomé ocho	/ _tomEOCo_ /
/ Eo /	tomé oscura	/ _tomEoskUra_ /	/ EU /	tomé única	/ _tomEUnika_ /
/ eu /	arte unido	/ _ArteunIDo_ /	/ Eu /	sé universitario	/ _sEuniBersitArjo_ /
/ E_ /	callé	/ _kaLE_ /	/ Ap /	tapia	/ _tApja_ /
/ at /	catar	/ _katAR_ /	/ Ak /	saca	/ _sAka_ /
/ am /	camión	/ _kamjOn_ /	/ an /	ganar	/ _ganAR_ /
/ aJ /	ensañarse	/ _ensaJArse_ /	/ aN /	hangar	/ _aNGAR_ /
/ AM /	ánfora	/ _AMfora_ /	/ AB /	cava	/ _kABa_ /
/ Af /	zafio	/ _TAfjo_ /	/ AT /	haz	/ _AT_ /
/ aD /	madera	/ _maDEra_ /	/ as /	ascenso	/ _asTEnso_ /
/ aC /	hachazo	/ _aCATo_ /	/ ax /	rajar	/ _RaxAR_ /
/ aG /	tragar	/ _traGAR_ /	/ al /	maldad	/ _malDAD_ /

Tabla B.10: Lista de difonemas y trifonemas (VIII).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ aL /	calló	/ _kaLO_ /	/ AR /	carro	/ _kARo_ /
/ ar /	tarea	/ _tarEa_ /	/ aj /	esta hiena	/ _EstajEna_ /
/ aw /	glaucoma	/ _glawkOma_ /	/ AI /	la isla	/ _lAIsla_ /
/ Ai /	la idea	/ _lAiDEa_ /	/ ai /	una idea	/ _UnaiDEa_ /
/ AE /	está ebrio	/ _estAEBrjo_ /	/ Ae /	está errado	/ _estAeRADo_ /
/ ae /	esa hermana	/ _EsaermAna_ /	/ AA /	está ágil	/ _estAAxil_ /
/ Aa /	está amable	/ _estAamABle_ /	/ aa /	esa amiga	/ _EsaamIGa_ /
/ Ao /	está hostil	/ _estAostIl_ /	/ AO /	está ogro	/ _estAOGro_ /
/ aO /	esta orden	/ _EstaOrDen_ /	/ AU /	está útil	/ _estAUtil_ /
/ Au /	la humana	/ _lAumAna_ /	/ au /	esa unidad	/ _EsauniDAD_ /
/ A_ /	allá	/ _aLA_ /	/ op /	opaco	/ _opAko_ /
/ ot /	motor	/ _motOR_ /	/ Ok /	roca	/ _ROka_ /
/ om /	comer	/ _komER_ /	/ on /	conocer	/ _konoTER_ /
/ oJ /	soñar	/ _soJAR_ /	/ oN /	roncar	/ _RoNkAR_ /
/ OM /	tacón fuerte	/ _takOMfwErte_ /	/ OB /	bobo	/ _bOBo_ /
/ Of /	cofre	/ _kOfre_ /	/ OT /	coz	/ _kOT_ /
/ oD /	odiar	/ _oDjAR_ /	/ os /	toser	/ _tosER_ /
/ oC /	ochenta	/ _oCEnta_ /	/ ox /	coger	/ _koxER_ /
/ OG /	ogro	/ _OGro_ /	/ Ol /	col	/ _kOl_ /
/ oL /	collar	/ _koLAR_ /	/ oR /	correr	/ _koRER_ /
/ or /	corazón	/ _koraTOn_ /	/ oj /	como hierro	/ _kOmojERo_ /
/ Ow /	comió huevos	/ _komjOwEBos_ /	/ OI /	tomó isla	/ _tomOIsla_ /
/ oi /	solo iré	/ _sOloirE_ /	/ Oi /	tomó Irán	/ _tomOirAn_ /
/ OE /	tomó esta	/ _tomOEsta_ /	/ oE /	miro esto	/ _mIroEsto_ /
/ oe /	cómo estar	/ _kOmoestAR_ /	/ Oa /	Lisboa	/ _lisBOa_ /
/ OA /	tomó alas	/ _tomOAlas_ /	/ oa /	miro allí	/ _mIroaLI_ /
/ OO /	tomó ocho	/ _tomOOCo_ /	/ Oo /	tomó objeto	/ _tomOoBxEto_ /
/ oo /	cojo objeto	/ _kOxooBxEto_ /	/ ou /	solo usted	/ _sOloustED_ /
/ Ou /	tomó usted	/ _tomOustED_ /	/ OU /	tomó uña	/ _tomOUJa_ /
/ O_ /	tomó	/ _tomO_ /	/ up /	agrupar	/ _aGrupAR_ /
/ ut /	brutal	/ _brutAl_ /	/ uk /	tucán	/ _tukAn_ /
/ um /	sumar	/ _sumAR_ /	/ un /	acunar	/ _akunAR_ /
/ uJ /	acuñar	/ _akuJAR_ /	/ uN /	ungir	/ _uNxIR_ /
/ uM /	álbum feo	/ _AlBuMfEo_ /	/ uB /	tubería	/ _tuBerIa_ /
/ Uf /	tufo	/ _tUfo_ /	/ uT /	bucear	/ _buTeAR_ /
/ uD /	dudar	/ _duDAR_ /	/ Us /	uso	/ _Uso_ /
/ uC /	escuchar	/ _eskuCAR_ /	/ Ux /	brujo	/ _brUxo_ /
/ UG /	Lugo	/ _lUGo_ /	/ Ul /	zulo	/ _TUlo_ /
/ ul /	culear	/ _kuleAR_ /	/ uL /	tullido	/ _tuLIDo_ /
/ uR /	zurrón	/ _TuROn_ /	/ ur /	curar	/ _kurAR_ /
/ uj /	tribu hiena	/ _trIBujEna_ /	/ uw /	tribu huérfana	/ _trIBuwErfana_ /
/ Ui /	su idiotez	/ _sUiDjotET_ /	/ uI /	espíritu indio	/ _espIrituInDjo_ /
/ ui /	espíritu indígena	/ _espIrituinDIxena_ /	/ UE /	su éxito	/ _sUEksito_ /
/ uE /	tribu épica	/ _trIBuEpika_ /	/ ue /	espíritu heredado	/ _espIrituereDADo_ /
/ Ua /	su amor	/ _sUamOR_ /	/ uA /	tribu aria	/ _trIBuArja_ /
/ ua /	tribu amable	/ _trIBuamABle_ /	/ Uo /	su olor	/ _sUolOR_ /
/ uO /	espíritu ocre	/ _espIrituOkre_ /	/ uo /	espíritu olvidado	/ _espIrituolBiDADo_ /
/ Uu /	su unión	/ _sUunjOn_ /	/ uU /	tribu única	/ _trIBuUnika_ /
/ uu /	tribu unida	/ _trIBuunIDa_ /	/ u_ /	tribu	/ _trIBu_ /
/ _i /	ideal	/ _iDeAl_ /	/ _E /	épico	/ _Epiko_ /
/ _A /	ábside	/ _ABsiDe_ /	/ _O /	ocio	/ _OTjo_ /
/ _U /	uña	/ _UJa_ /	/ pli /	cómplice	/ _kOmpliTe_ /
/ ple /	plegar	/ _pleGAR_ /	/ pla /	aplanar	/ _aplanAR_ /
/ plo /	explosivo	/ _eksplosIBo_ /	/ plu /	plumilla	/ _plumILa_ /
/ pri /	primero	/ _primEro_ /	/ pre /	prestar	/ _prestAR_ /
/ pra /	practicar	/ _praktikAR_ /	/ prO /	propio	/ _prOpjo_ /
/ prU /	Prusia	/ _prUsja_ /	/ pje /	piedad	/ _pjeDAD_ /
/ pja /	pianola	/ _pjanOla_ /	/ pjo /	apio	/ _Apjo_ /
/ pwe /	pueblerino	/ _pweBlerIno_ /	/ pwa /	puaré	/ _pwarE_ /
/ bli /	blincar	/ _bliNkAR_ /	/ ble /	blefaritis	/ _blefarItis_ /
/ bla /	blancura	/ _blaNkUra_ /	/ blo /	bloquear	/ _blokeAR_ /
/ blu /	blusón	/ _blusOn_ /	/ bri /	brigada	/ _briGADa_ /
/ brE /	brecha	/ _brECa_ /	/ bra /	bracear	/ _braTeAR_ /
/ bro /	bromear	/ _bromeAR_ /	/ bru /	brusquedad	/ _bruskeDAD_ /

Tabla B.11: Lista de difonemas y trifonemas (IX).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ bje /	vienés	/ _bjenEs_ /	/ bja /	viajar	/ _bjaxAR_ /
/ bjo /	violar	/ _bjolAR_ /	/ bwi /	Buitrago	/ _bwitrAGo_ /
/ bwe /	buenaventura	/ _bwenaBentUra_ /	/ bwa /	buhardilla	/ _bwarDILa_ /
/ bwO /	Buol	/ _bwOl_ /	/ trI /	trigo	/ _trIGo_ /
/ tre /	trenzar	/ _trenTAR_ /	/ tra /	tratado	/ _tratADo_ /
/ trO /	tronco	/ _trONko_ /	/ tru /	trufar	/ _trufAR_ /
/ tje /	tiendecita	/ _tjenDeTIta_ /	/ tja /	tialina	/ _tjalIna_ /
/ tjO /	catión	/ _katjOn_ /	/ twi /	fatuidad	/ _fatwiDAD_ /
/ twe /	tuerquecilla	/ _twerkeTILa_ /	/ twA /	perpetuar	/ _perpetwAR_ /
/ twO /	perpetuó	/ _perpetwO_ /	/ drI /	dril	/ _drIl_ /
/ drE /	Dresde	/ _drEsDe_ /	/ dra /	dragón	/ _draGOn_ /
/ drO /	droga	/ _drOGa_ /	/ dje /	dietética	/ _djetEtika_ /
/ dja /	diarrea	/ _djaREa_ /	/ djo /	Dionisio	/ _djonIsjo_ /
/ dwe /	duendecillo	/ _dwenDeTILo_ /	/ dwA /	dual	/ _dwAl_ /
/ kli /	climatizado	/ _klimatiTADo_ /	/ klE /	clero	/ _klEro_ /
/ kla /	aclarar	/ _aklarAR_ /	/ klo /	clonar	/ _klonAR_ /
/ klu /	reclutar	/ _ReklutAR_ /	/ krI /	Cristo	/ _krIsto_ /
/ kre /	Cretáceo	/ _kretATeo_ /	/ kra /	craneal	/ _kraneAl_ /
/ kro /	cromado	/ _kromADo_ /	/ kru /	cruzar	/ _kruTAR_ /
/ kje /	quietud	/ _kjetUD_ /	/ kjA /	esquiar	/ _eskjAR_ /
/ kjO /	kiosco	/ _kjOsko_ /	/ kwI /	cuido	/ _kwIDo_ /
/ kwe /	cuentista	/ _kwentIsta_ /	/ kwA /	cuánto	/ _kwAnto_ /
/ kwO /	cuota	/ _kwOta_ /	/ glI /	glíptica	/ _glIptika_ /
/ gle /	glebita	/ _gleBIta_ /	/ glA /	glándula	/ _glAnDula_ /
/ glo /	global	/ _gloBAl_ /	/ glu /	glucosa	/ _glukOsa_ /
/ gri /	grisáceo	/ _grisATeo_ /	/ grE /	gremio	/ _grEmjo_ /
/ gra /	grabar	/ _graBAR_ /	/ grO /	groso	/ _grOso_ /
/ gru /	gruñir	/ _gruJIR_ /	/ gja /	guiaré	/ _gjarE_ /
/ gjo /	guionista	/ _gjonIsta_ /	/ gwe /	güemul	/ _gwemUl_ /
/ gwa /	guantazo	/ _gwantATo_ /	/ gwi /	güisquería	/ _gwiskerIa_ /
/ gwI /	güisqui	/ _gwIski_ /	/ mje /	miedoso	/ _mjeDOso_ /
/ mjA /	mialgia	/ _mjAlxja_ /	/ mjO /	miope	/ _mjOpe_ /
/ mwe /	muestreo	/ _mwestrEo_ /	/ mwA /	ermuarra	/ _ermwARa_ /
/ nje /	nietecito	/ _njeteTIto_ /	/ nja /	Alemania	/ _alemAnja_ /
/ njo /	aluminio	/ _alumInjo_ /	/ nwe /	tenue	/ _tEnwe_ /
/ nwa /	manualidad	/ _manwaliDAD_ /	/ nwo /	continuo	/ _kontInwo_ /
/ BlI /	oblicuo	/ _oBlIkwo_ /	/ BlE /	hablemos	/ _aBlEmos_ /
/ Bla /	hablaré	/ _aBlarE_ /	/ Blo /	hablo	/ _ABlo_ /
/ BlU /	la blusa	/ _lABlUsa_ /	/ Bri /	abrigar	/ _aBriGAR_ /
/ BrE /	cobré	/ _koBrE_ /	/ BrA /	habrá	/ _aBrA_ /
/ BrO /	abróchate	/ _aBrOCate_ /	/ BrU /	abrúmate	/ _aBrUmate_ /
/ Bje /	cambie	/ _kAmBje_ /	/ BjA /	cambiar	/ _kamBjAR_ /
/ Bjo /	cambio	/ _kAmBjo_ /	/ Bwe /	abuelito	/ _aBwelIto_ /
/ BwI /	cinco buitres	/ _TINkoBwItres_ /	/ Bwi /	de Buitrago	/ _dEBwitrAGo_ /
/ Bwo /	el buhonero	/ _ElBwonEro_ /	/ BwO /	de Buol	/ _dEBwOl_ /
/ Bwa /	la buhardilla	/ _lABwarDILa_ /	/ BwA /	sí buana	/ _sIBwAna_ /
/ flI /	aflígete	/ _aflIxete_ /	/ fle /	flechazo	/ _fleCATo_ /
/ fla /	flaquear	/ _flakeAR_ /	/ flo /	florecer	/ _floreTER_ /
/ flu /	fluvial	/ _fluBjAl_ /	/ fri /	frigorífico	/ _friGorIfiko_ /
/ fre /	fresón	/ _fresOn_ /	/ fra /	fraterno	/ _fratErno_ /
/ frO /	frontis	/ _frOntis_ /	/ fru /	afrutado	/ _afrutADo_ /
/ fje /	fiereza	/ _fjerETa_ /	/ fja /	fiambrera	/ _fjamBrEra_ /
/ fjO /	Fiona	/ _fjOna_ /	/ fwe /	Fuensanta	/ _fwensAnta_ /
/ fwA /	fuá	/ _fwA_ /	/ Tje /	ciempiés	/ _TjempjEs_ /
/ Tja /	cianuro	/ _TjanUro_ /	/ Tjo /	recio	/ _RETjo_ /
/ Twe /	Pozuelano	/ _poTwelAno_ /	/ Dri /	apadrinar	/ _apaDrinAR_ /
/ DrE /	tendré	/ _tenDrE_ /	/ DrA /	tendrá	/ _tenDrA_ /
/ DrO /	padrón	/ _paDrOn_ /	/ Dru /	madrugar	/ _maDruGAR_ /
/ DjE /	adiestro	/ _aDjEstro_ /	/ DjA /	radiar	/ _RaDjAR_ /
/ Djo /	endiosar	/ _enDjosAR_ /	/ DwI /	beduino	/ _beDwIno_ /
/ Dwi /	balduinista	/ _balDwinIsta_ /	/ DwE /	el duende	/ _ElDwEnDe_ /
/ DwA /	gradual	/ _graDwAl_ /	/ DwO /	arduoso	/ _arDwOso_ /
/ sje /	sienita	/ _sjenIta_ /	/ sja /	Asia	/ _Asja_ /
/ sjO /	pasión	/ _pasjOn_ /	/ swi /	suicida	/ _swiTIDa_ /

Tabla B.12: Lista de difonemas y trifonemas (X).

Unidad	Palabra	Transcripción	Unidad	Palabra	Transcripción
/ swe /	sueldazo	/ _swelDATo_ /	/ swa /	suavizante	/ _swaBiTAnte_ /
/ xje /	jienense	/ _xjenEnse_ /	/ xja /	regia	/ _RExja_ /
/ xjo /	regio	/ _RExjo_ /	/ xwi /	enjuiciar	/ _eNxwiTjAR_ /
/ xwe /	juerguista	/ _xwerGIsta_ /	/ xwa /	juanete	/ _xwanEte_ /
/ Gle /	ingle	/ _INGle_ /	/ GlA /	seglar	/ _seGlAR_ /
/ GlO /	renglón	/ _ReNGlOn_ /	/ Glu /	aglutinar	/ _aGlutinAR_ /
/ GrI /	agrícola	/ _aGrIkola_ /	/ Gre /	agresión	/ _aGresjOn_ /
/ Gra /	agravar	/ _aGraBAR_ /	/ Gro /	agropecuario	/ _aGropekwArjo_ /
/ GrU /	agrúpate	/ _aGrUpate_ /	/ GjE /	siguiente	/ _siGjEnte_ /
/ Gja /	le guiará	/ _lEGjarA_ /	/ Gjo /	siete guionistas	/ _sjEteGjonIstas_ /
/ GwE /	halagüeño	/ _alaGwEJo_ /	/ GwA /	antigualla	/ _antiGwALa_ /
/ GwO /	aguó la fiesta	/ _aGwOlAfjEsta_ /	/ GwI /	argüir	/ _arGwIR_ /
/ Gwi /	argüirá	/ _arGwirA_ /	/ ljE /	liebre	/ _ljEBre_ /
/ lja /	camelia	/ _kamElja_ /	/ ljo /	folio	/ _fOljo_ /
/ lwe /	hasta lueguito	/ _AstalweGIto_ /	/ lwa /	evaluaremos	/ _eBalwarEmos_ /
/ lwo /	superfluo	/ _supErflwo_ /	/ lwi /	Luisito	/ _lwisIto_ /
/ Lwe /	pilluelito	/ _piLwelIto_ /	/ Rje /	arriesgado	/ _aRjesGADo_ /
/ Rja /	arriarás	/ _aRjarAs_ /	/ Rjo /	arriostrado	/ _aRjostrADo_ /
/ Rwi /	ruidoso	/ _RwiDOso_ /	/ Rwe /	ruedero	/ _RweDEro_ /
/ Rwa /	ruanés	/ _RwanEs_ /	/ rje /	aries	/ _Arjes_ /
/ rjA /	variable	/ _barjABle_ /	/ rjO /	Iriondo	/ _irjOnDo_ /
/ prA /	práctica	/ _prAktika_ /	/ Tju /	ciudadano	/ _TjuDaDAno_ /
/ TjU /	Ciuro	/ _TjUro_ /	/ njU /	Niurca	/ _njUrka_ /
/ nju /	Niubó	/ _njuBO_ /	/ CA /	enganchar	/ _eNGanCAR_ /
/ Ci /	archivar	/ _arCiBAR_ /	/ CE /	ochenta	/ _oCEnta_ /
/ EN /	luengo	/ _lwENGo_ /	/ eT /	ensordecer	/ _ensorDeTER_ /
/ iB /	adivino	/ _aDiBIno_ /	/ If /	frigorífico	/ _friGorIfiko_ /
/ IT /	barniz	/ _barnIT_ /	/ iw /	orihuela	/ _oriwEla_ /
/ m_ /	drum	/ _drUm_ /	/ pT /	concepción	/ _konTepTjOn_ /
/ rw /	teruel	/ _terwEl_ /	/ TG /	es juzgado	/ _EsxuTGADo_ /
/ ew /	reunión	/ _RewnjOn_ /	/ kn /	arácnido	/ _arAkniDo_ /
/ Cu /	chubascos	/ _CuBAskos_ /	/ Gm /	Segmento	/ _seGmEnto_ /
/ wC /	Agauchar	/ _aGawCAR_ /	/ ii /	Api inicial	/ _ApiiniTjAl_ /
/ io /	Api oscura	/ _ApioskUra_ /	/ dru /	drupáceo	/ _drupATeo_ /
/ fwi /	Fuitá	/ _fwitA_ /	/ Cja /	Salvachia	/ _salBACja_ /
/ ad /	adxós	/ _adksOs_ /	/ Dl /	Adlátere	/ _aDlAtere_ /
/ t_ /	Tarot	/ _tarOt_ /	/ k_ /	Toc	/ _tOk_ /
/ kjU /	Desquiú	/ _deskjU_ /	/ kp /	Tic parado	/ _tIkparADo_ /
/ kx /	Tic genial	/ _tIkxenjAl_ /	/ L_ /	Coll	/ _kOL_ /
/ mD /	Sam duro	/ _sAmDUro_ /	/ mk /	Sam cuñado	/ _sAmkuJADo_ /
/ ml /	Kremlin	/ _krEmlin_ /	/ ms /	Módems	/ _mODems_ /
/ p_ /	Chip	/ _CIp_ /	/ pB /	Top vacío	/ _tOpBaTIo_ /
/ pf /	Cap flojo	/ _kApflOxo_ /	/ ps /	Necropsia	/ _nekrOpsja_ /
/ sju /	Siujar	/ _sjuxAR_ /	/ tjU /	Veintiuno	/ _bejntjUno_ /
/ tk /	Chat corto	/ _CAtkOrto_ /	/ tp /	Postpalatal	/ _postpalatAl_ /
/ ts /	Robots	/ _RoBOts_ /	/ w_ /	Tau	/ _tAw_ /
/ xjO /	surgió	/ _surxjO_ /	/ nu /	nubarrón	/ _nuBaROn_ /
/ xi /	reloj inglés	/ _RelOxiNGlEs_ /	/ GB /	rugby	/ _RUGBi_ /
/ gT /	zigzag	/ _TigTAG_ /	/ Djo /	Odio	/ _ODjo_ /

Apéndice C

Análisis estadístico de los parámetros prosódicos del corpus

C.1. Duración segmental

En este apartado se muestran los resultados del análisis de la media y de la desviación típica de las duraciones de cada fonema del corpus en función del estilo y, también, para el conjunto del corpus (véanse las tablas C.1, C.2 y C.3).

Tabla C.1: Duración media, desviación estándar y frecuencia absoluta de aparición de los segmentos del corpus en los estilos neutro y alegre

	NEUTRO				ALEGRE		
Fon.	**μ (ms)**	**σ (ms)**	**Núm**	**Fon.**	**μ (ms)**	**σ (ms)**	**Núm**
a	82	34	1816	a	108	49	1885
A	85	28	1409	A	109	36	1608
e	68	32	1664	e	85	39	1692
E	66	28	1892	E	91	43	2045
i	64	21	614	i	76	26	772
I	90	32	665	I	108	35	691
o	79	34	1442	o	100	50	1496
O	78	31	1015	O	106	40	920
u	67	19	161	u	73	21	221
U	79	29	487	U	98	32	449
j	78	29	682	j	79	32	692
w	68	30	230	w	65	27	258
p	85	27	676	p	87	27	692
t	78	25	1316	t	76	22	1317
k	81	28	1119	k	82	25	1009
b	137	45	57	b	111	31	27
B	53	19	636	B	51	18	660
d	102	39	58	d	97	32	186
D	49	19	1066	D	51	19	1181
g	122	48	7	g	147	42	42
G	47	17	201	G	60	23	274
n	75	33	1678	n	84	36	1670
m	77	28	922	m	89	32	876
J	103	20	59	J	102	19	77
s	100	37	2050	s	104	40	2095
x	122	24	183	x	114	22	246
C	131	20	51	C	124	31	39
T	108	28	537	T	97	26	479
r	48	15	1438	r	46	20	1626
R	104	33	196	R	115	46	246
l	67	27	1149	l	80	37	1524
L	95	40	72	L	103	35	74
f	100	29	211	f	94	31	201
N	81	21	126	N	88	28	116
M	71	15	40	M	82	37	35
Sil	334	179	1508	Sil	211	104	1511

Tabla C.2: Duración media, desviación estándar y frecuencia absoluta de aparición de los segmentos del corpus en los estilos sensual y agresivo

	SENSUAL				AGRESIVO		
Fon.	μ **(ms)**	σ **(ms)**	**Núm**	**Fon.**	μ **(ms)**	σ **(ms)**	**Núm**
a	90	39	1590	a	92	43	2621
A	97	34	1064	A	102	37	2315
e	81	34	1200	e	74	36	2697
E	82	36	1413	E	74	32	2974
i	72	25	519	i	78	28	853
I	101	34	559	I	115	39	976
o	86	38	1016	o	102	54	2391
O	89	36	781	O	107	42	1647
u	65	20	148	u	78	24	299
U	86	31	446	U	115	56	851
j	92	30	595	j	90	32	1047
w	74	28	194	w	86	29	403
p	107	35	656	p	85	34	1130
t	90	33	995	t	77	30	1919
k	105	35	729	k	76	28	1652
b	135	45	48	b	115	33	56
B	58	19	474	B	48	19	1005
d	109	46	70	d	88	31	93
D	58	22	831	D	44	18	1735
g	111	18	6	g	105	56	18
G	58	22	191	G	51	23	412
n	90	34	1262	n	78	39	2579
m	87	35	728	m	80	32	1379
J	116	26	35	J	111	21	83
s	122	50	1337	s	103	37	2961
x	140	31	179	x	101	28	362
C	157	33	33	C	120	24	166
T	131	35	416	T	87	32	690
r	57	23	1223	r	45	18	2321
R	121	54	246	R	90	41	310
l	80	35	1117	l	72	29	1924
L	95	29	83	L	117	30	120
f	123	37	242	f	91	35	313
N	92	36	77	N	74	34	264
M	78	23	26	M	74	26	36
Sil	321	157	1465	Sil	434	457	1698

Tabla C.3: Duración media, desviación estándar y frecuencia absoluta de aparición de los segmentos del corpus en el estilo triste y en el conjunto del corpus

	TRISTE				TOTAL		
Fon.	μ **(ms)**	σ **(ms)**	**Núm.**	**Fon.**	μ **(ms)**	σ **(ms)**	**Núm.**
a	93	48	2467	a	93	43	10379
A	98	39	1918	A	99	35	8314
e	77	41	1976	e	77	37	9229
E	74	37	2347	E	77	35	10671
i	90	30	599	i	76	26	3357
I	132	56	815	I	111	40	3706
o	103	63	1746	o	96	50	8091
O	97	48	1099	O	97	40	5462
u	83	30	270	u	75	24	1099
U	103	54	694	U	99	44	2927
j	99	35	843	j	88	32	3859
w	87	36	311	w	78	30	1396
p	89	57	805	p	90	36	3959
t	80	47	1527	t	80	32	7074
k	91	49	1119	k	85	33	5628
b	136	68	141	b	131	52	329
B	58	21	806	B	53	19	3581
d	114	65	160	d	102	44	567
D	56	37	1355	D	51	23	6168
g	118	50	27	g	128	46	100
G	55	37	274	G	54	25	1352
n	99	48	2009	n	85	39	9198
m	106	37	1147	m	88	33	5052
J	132	32	73	J	113	23	327
s	103	38	2568	s	105	39	11011
x	133	29	282	x	120	27	1252
C	142	24	65	C	129	25	354
T	115	28	564	T	106	30	2686
r	65	21	1779	r	52	19	8387
R	111	51	357	R	108	46	1355
l	91	45	1604	l	78	35	7318
L	119	49	118	L	108	37	467
f	106	34	221	f	102	33	1188
N	111	41	125	N	86	32	708
M	98	17	25	M	79	24	162
Sil	650	329	1553	Sil	393	251	7735

C.2. Frecuencia fundamental

En las figuras siguientes se muestra la distribución de la media de F_0 en cada estilo en función de diferentes atributos prosódicos utilizados por el sistema de modelado y predicción de la melodía.

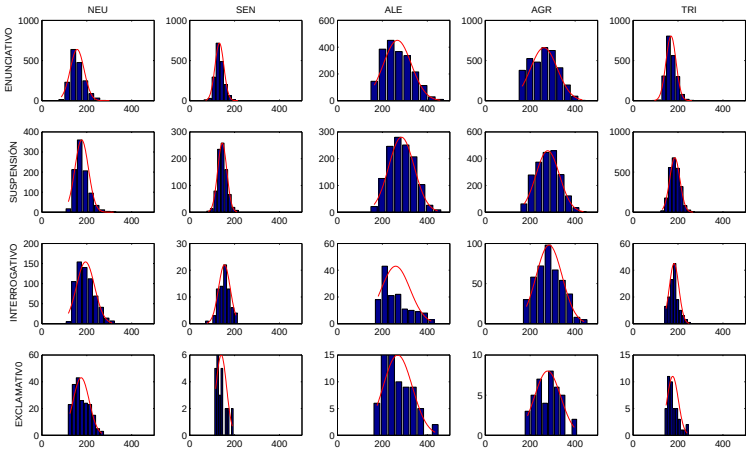


Figura C.1: Distribución de la media de F_0 en función del atributo TIPO-GE en cada estilo

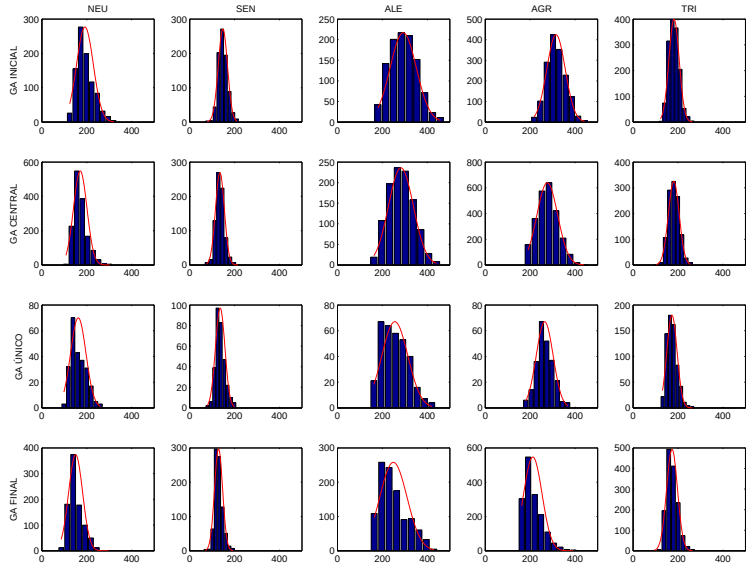


Figura C.2: Distribución de la media de F_0 en función del atributo GA-en-GE en cada estilo

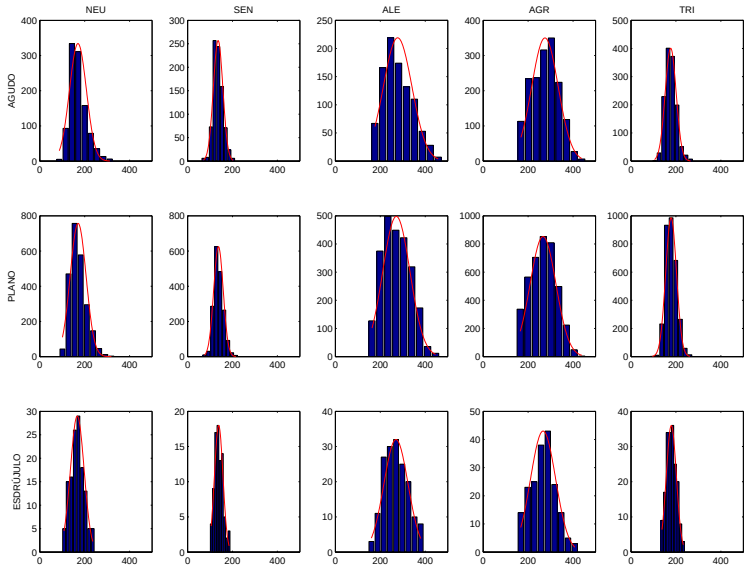


Figura C.3: Distribución de la media de F_0 en función del atributo ACENTO en cada estilo

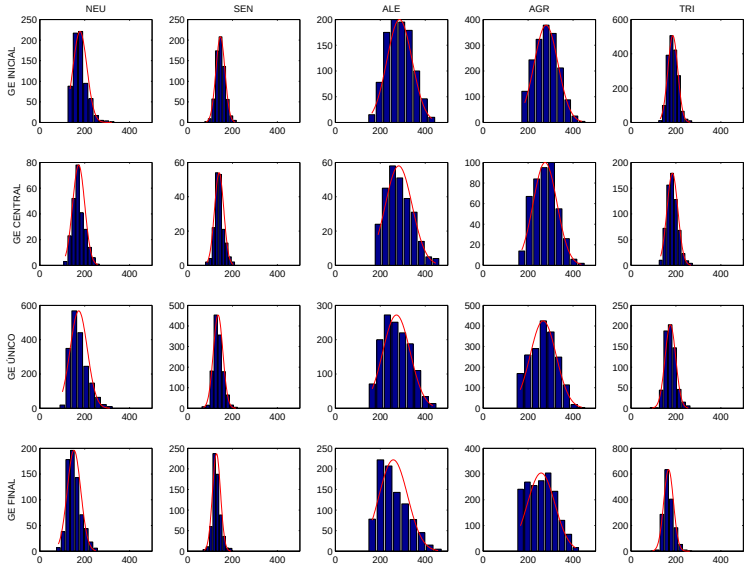


Figura C.4: Distribución de la media de F_0 en función del atributo GA-en-FRA en cada estilo

Apéndice D

Prueba subjetiva para la evaluación del modelado prosódico

En este anexo, se reproducen las frases utilizadas en la prueba subjetiva de evaluación del módulo de generación automática de los parámetros prosódicos; se detallan los valores de raíz del error cuadrático medio —*Root Mean Squared Error*— (RMSE) y de coeficiente de correlación de Pearson (ρ) obtenidos en los tres parámetros estimados y, finalmente, se representan las gráficas a partir de los valores reales y de los estimados.

D.1. Estilo neutro

Las frases escogidas para la prueba subjetiva de evaluación del módulo de predicción de los parámetros prosódicos en el estilo neutro son las siguientes:

1. Ocho jóvenes actores, acabarán colgados en tu habitación.
2. A partir de ahora sus empleados van a hablar por tres.
3. Absoluta perfección mecánica.
4. Ahora cuesta aún menos hablar con los que están lejos.
5. Antes de acudir al psicólogo, visite su quiosco.
6. Bienvenidos al futuro de la tercera generación de móviles.
7. La primera red privada de comunicaciones, para pymes.
8. Perfiles para ganar en planificación.
9. ¡El medio de publicidad más rentable de la región!
10. ¡Trescientos kilómetros por hora!
11. ¡El mejor monitor de su ordenador!
12. ¿Creía saberlo todo sobre márqueting financiero?
13. Ninguna imprime más rápido.
14. ¿Necesitas algo más para convencerte?
15. ¿No es increíble lo que hace una buena programación?

Tabla D.1: Valores promedio de RMSE y de ρ para los tres parámetros prosódicos de las frases que forman la prueba subjetiva en el estilo neutro.

	F0 (Hz)		Duración (ms)		Energía (rms)	
Frase	RMSE	ρ	RMSE	ρ	RMSE	ρ
1	39.11	0.38	19.71	0.78	0.027	0.75
2	31.13	0.66	21.27	0.82	0.022	0.75
3	28.54	0.69	20.50	0.75	0.015	0.91
4	31.26	0.46	22.12	0.72	0.030	0.85
5	24.18	0.82	18.40	0.79	0.028	0.74
6	24.41	0.63	18.32	0.86	0.015	0.87
7	37.45	0.47	17.60	0.80	0.022	0.78
8	23.04	0.64	15.41	0.88	0.032	0.88
9	23.89	0.47	23.04	0.69	0.021	0.80
10	31.07	0.65	18.40	0.82	0.015	0.90
11	23.28	0.76	20.21	0.65	0.020	0.85
12	37.38	0.55	18.68	0.70	0.041	0.66
13	31.37	0.57	27.12	0.57	0.033	0.73
14	38.33	0.60	17.91	0.82	0.017	0.90
15	37.02	0.68	20.10	0.69	0.026	0.74

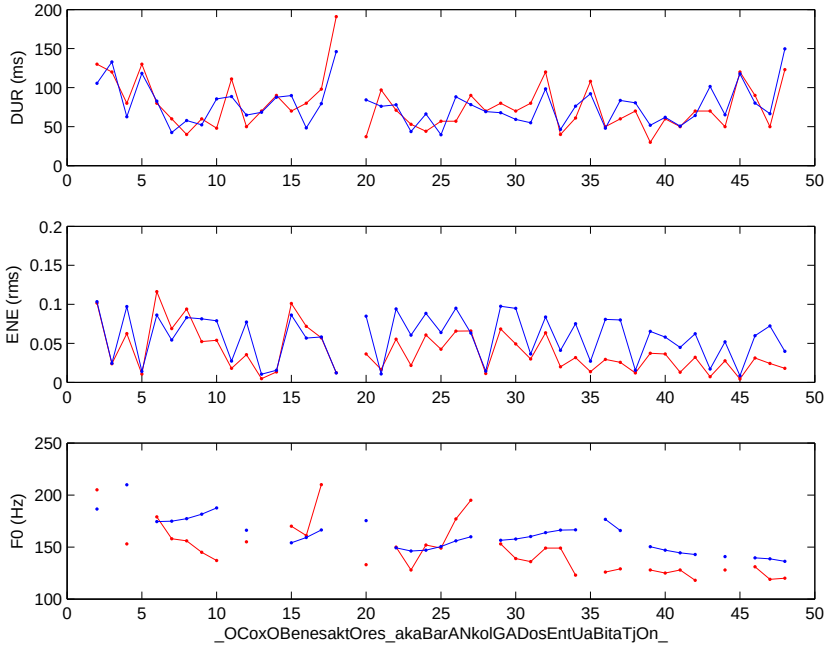


Figura D.1: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 1 del estilo neutro.

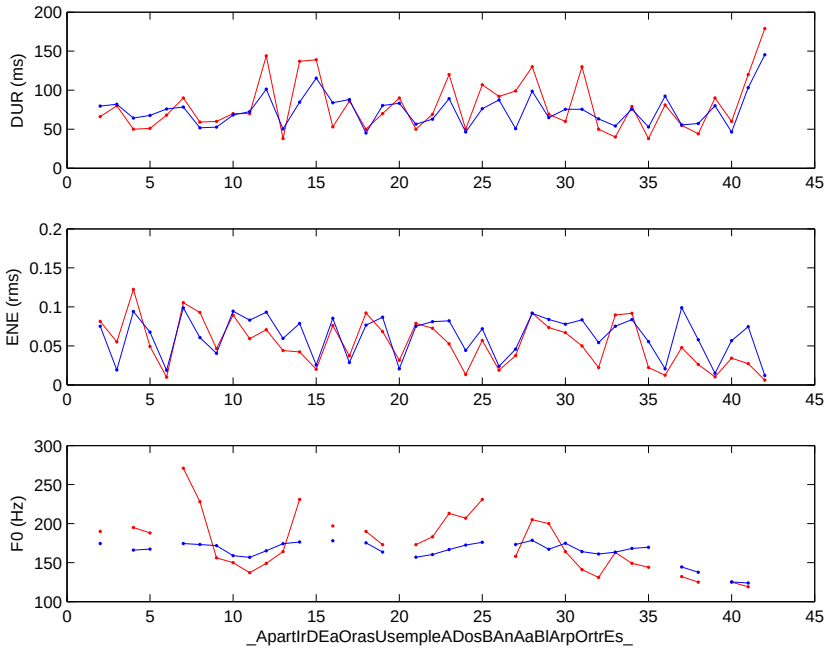


Figura D.2: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 2 del estilo neutro.

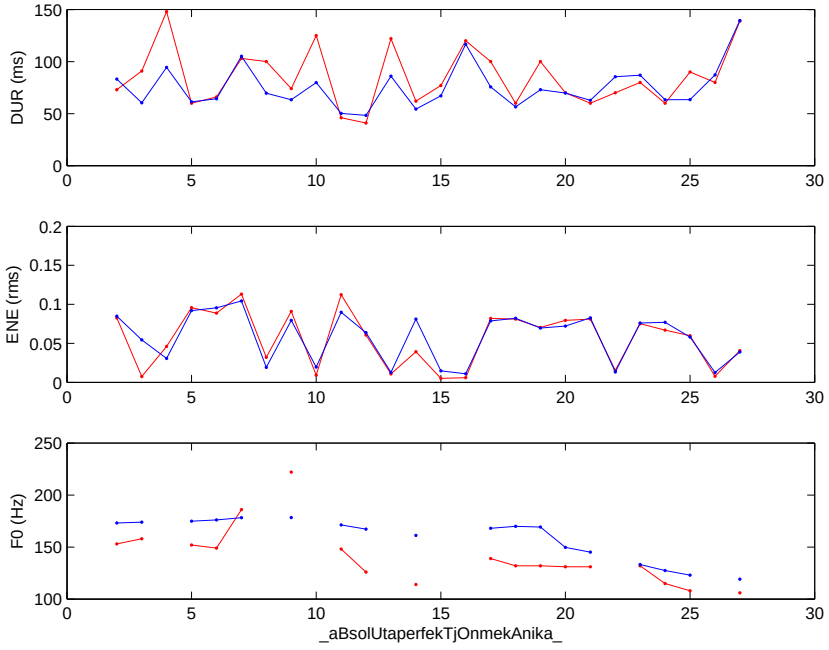


Figura D.3: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 3 del estilo neutro.

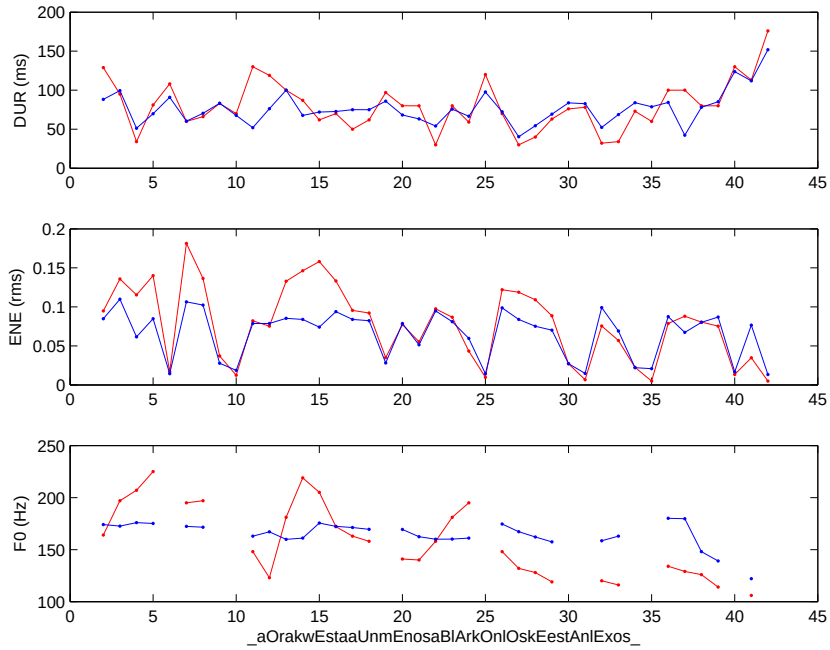


Figura D.4: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 4 del estilo neutro.

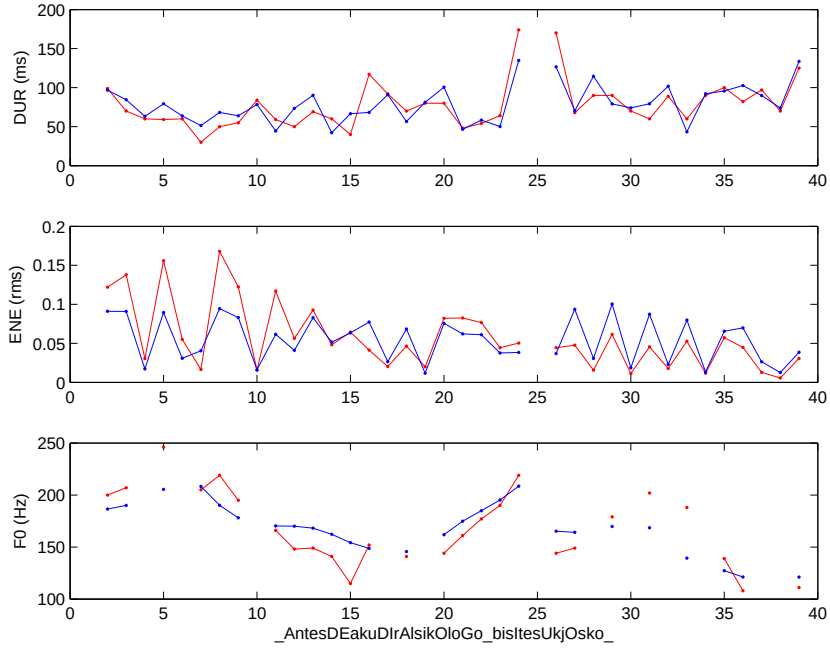


Figura D.5: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 5 del estilo neutro.

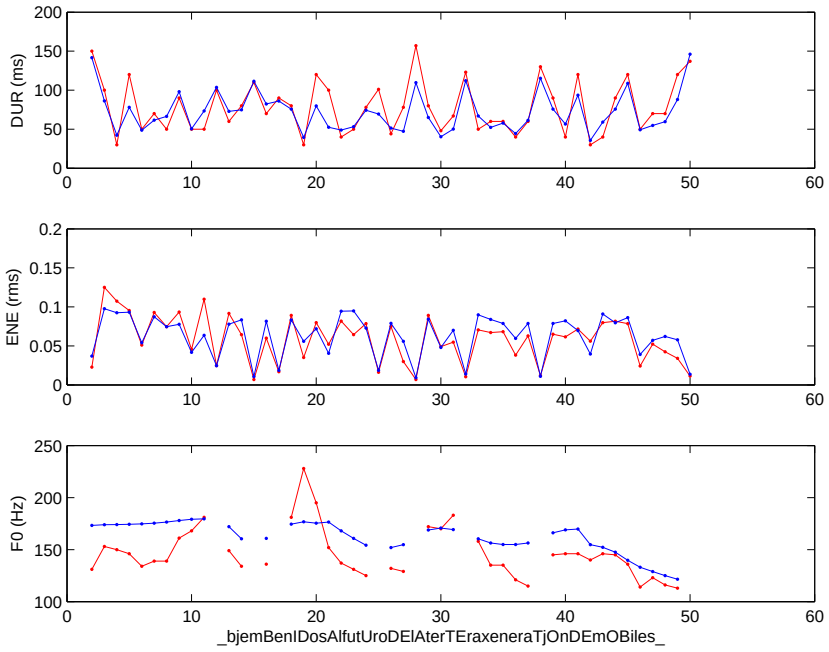


Figura D.6: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 6 del estilo neutro.

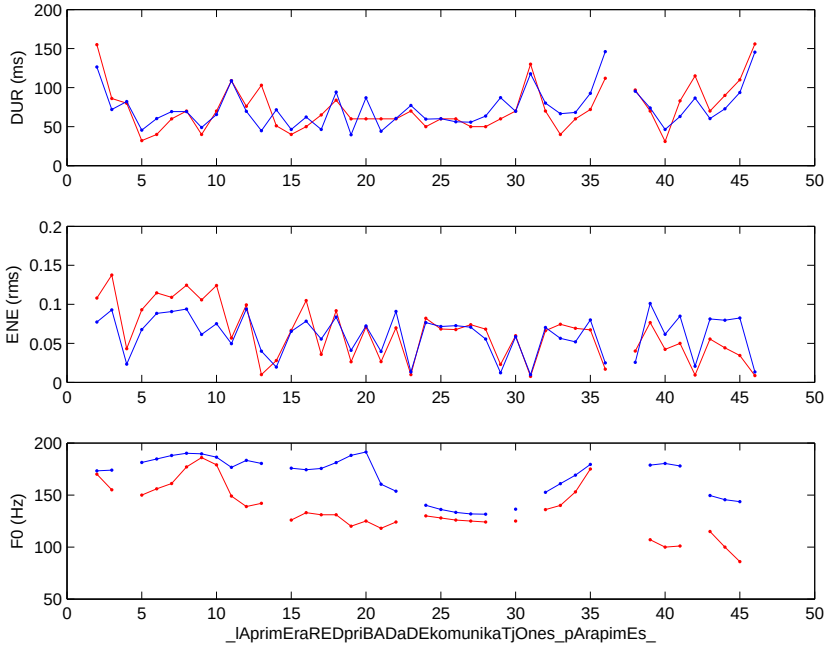


Figura D.7: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 7 del estilo neutro.

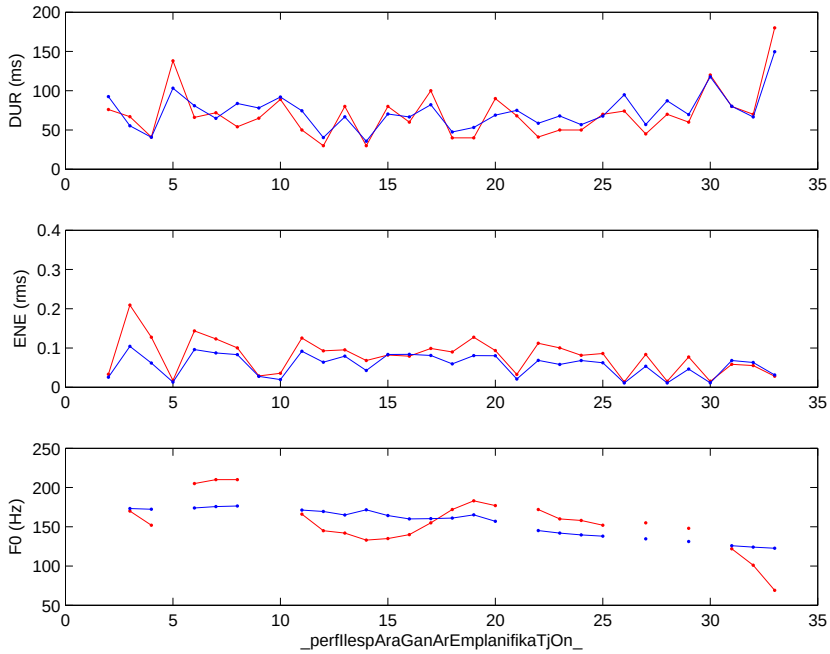


Figura D.8: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 8 del estilo neutro.

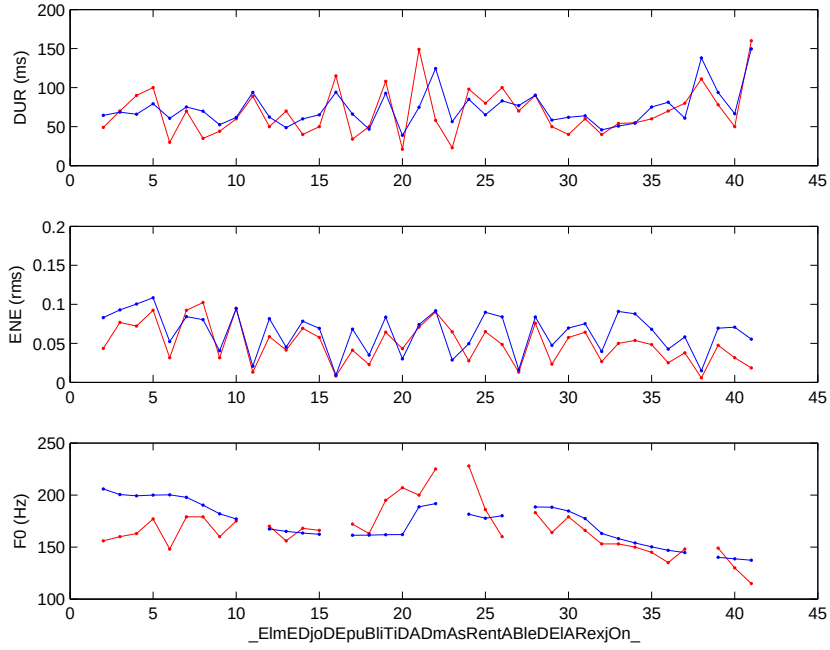


Figura D.9: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 9 del estilo neutro.

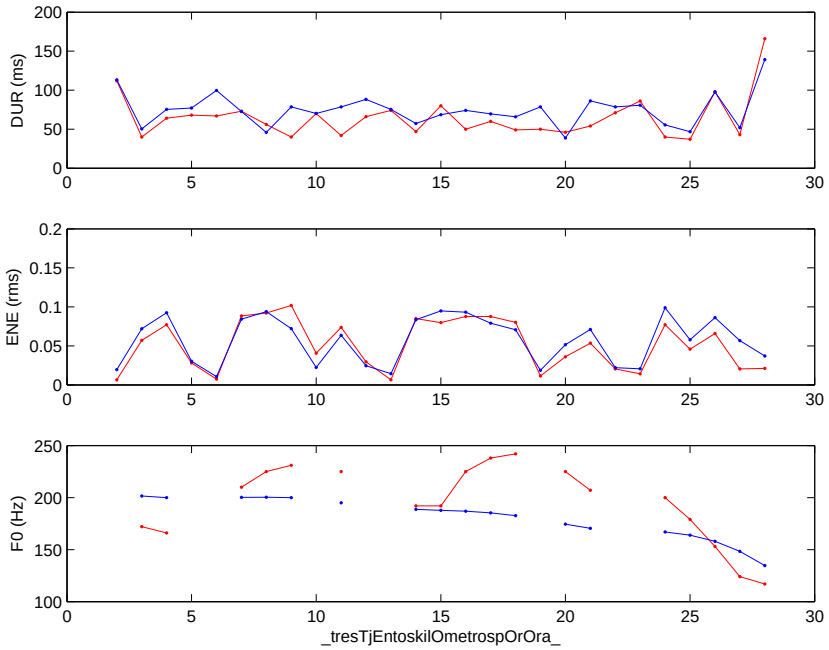


Figura D.10: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 10 del estilo neutro.

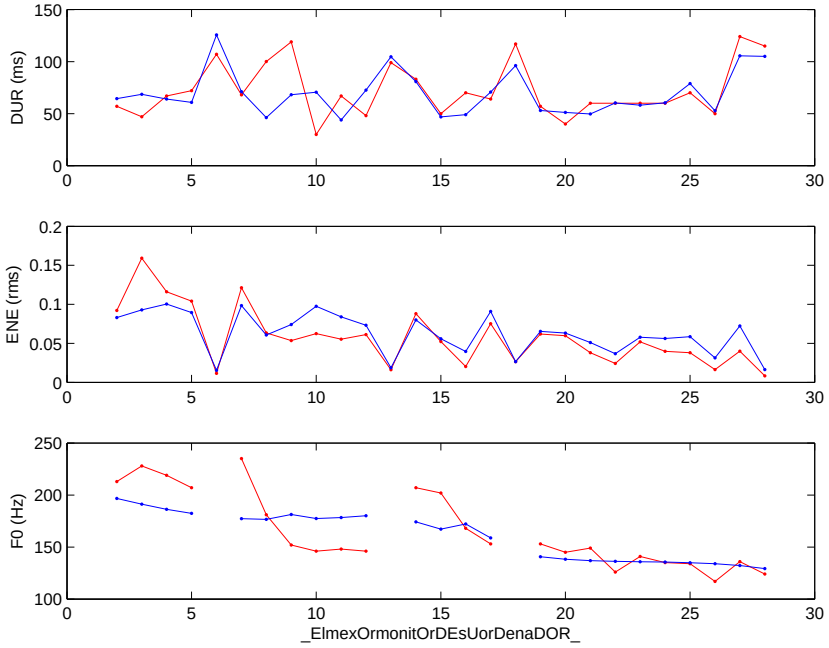


Figura D.11: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 11 del estilo neutro.

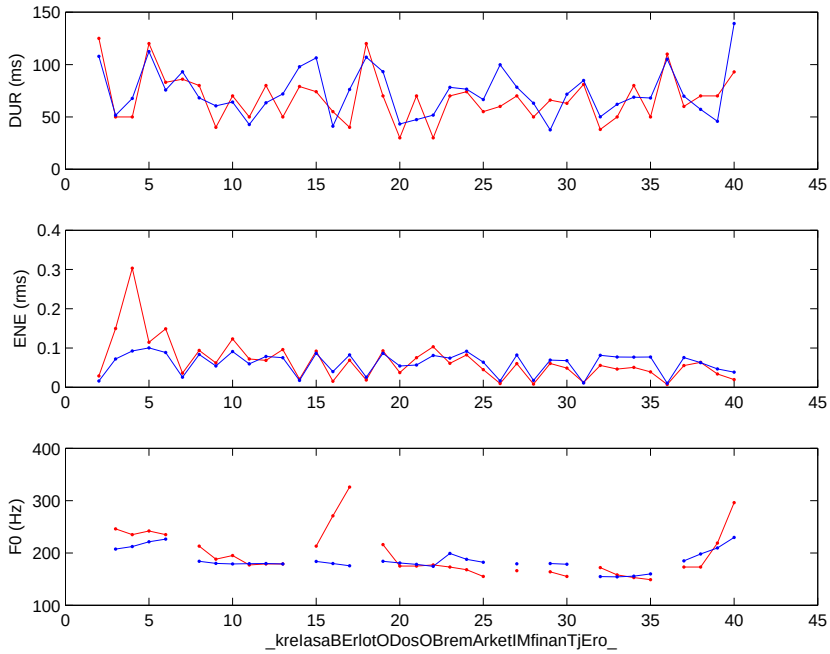


Figura D.12: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 12 del estilo neutro.

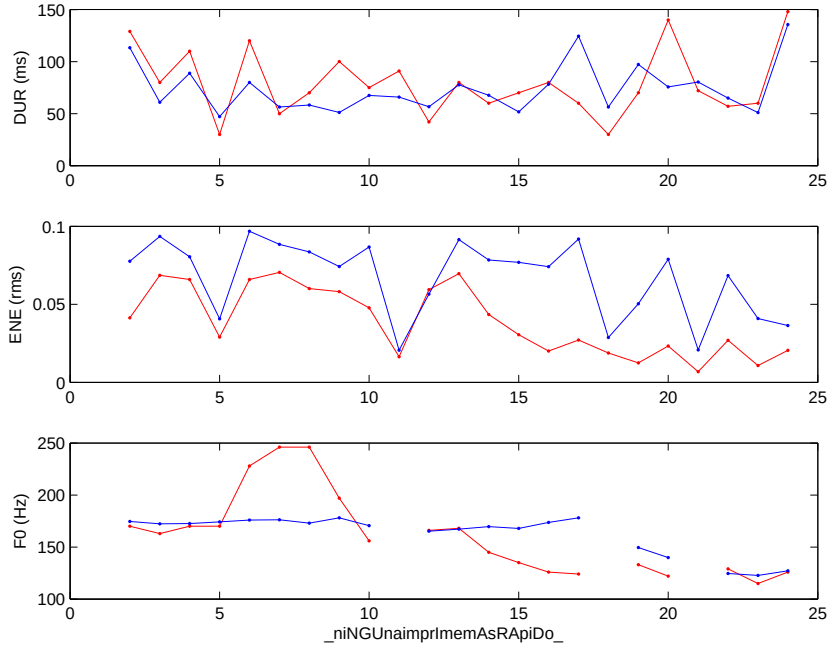


Figura D.13: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 13 del estilo neutro.

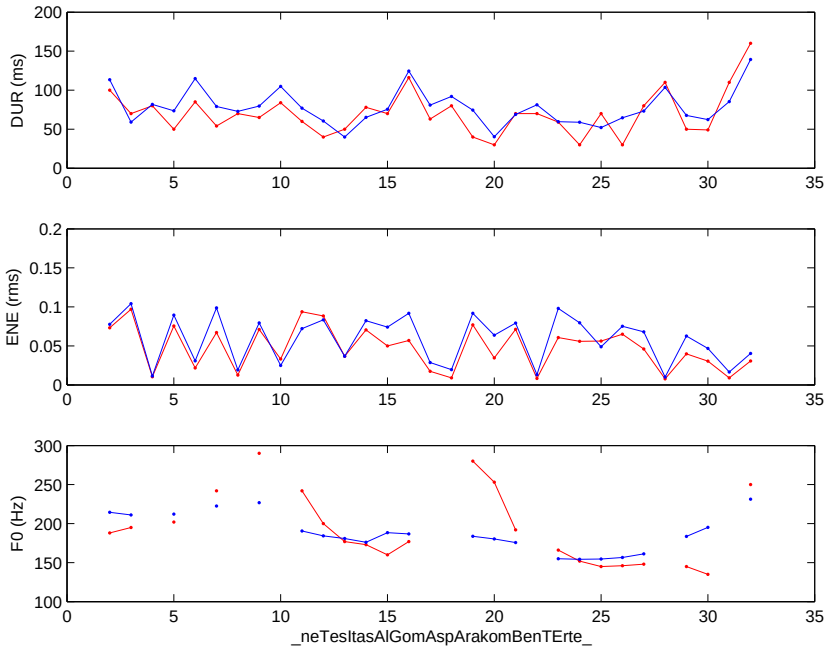


Figura D.14: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 14 del estilo neutro.

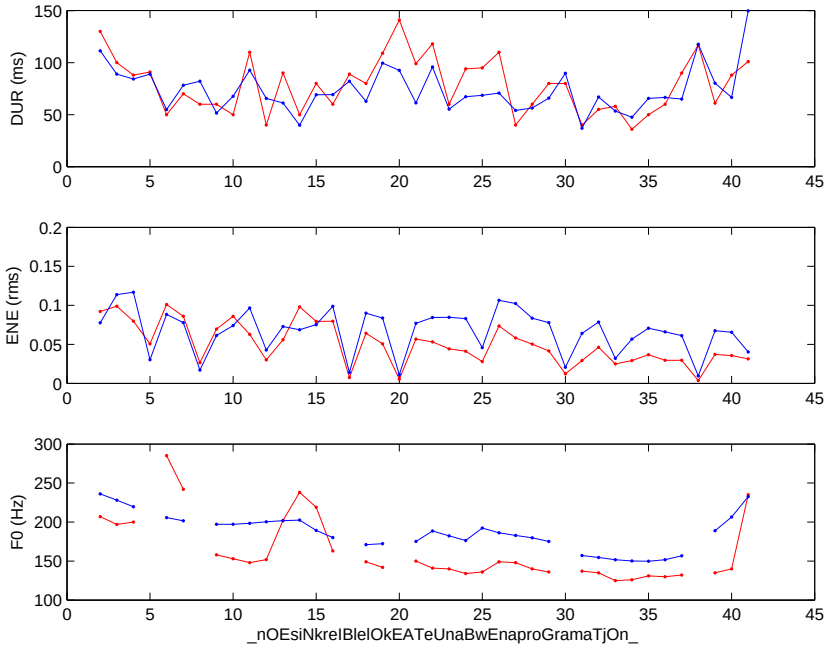


Figura D.15: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 15 del estilo neutro.

D.2. Estilo sensual

Las frases escogidas para la prueba subjetiva de evaluación del módulo de predicción de los parámetros prosódicos en el estilo sensual son las siguientes:

1. ¿Hasta qué punto aprecias tus miembros?
2. Porque no hay dos pieles iguales.
3. Tu tratamiento completo, de regalo.
4. Colección, primavera verano, dos mil.
5. Hemos mejorado nuestra mayor protección.
6. Igual que andar por la arena de la playa.
7. La proeza del color de larga duración, confortable y ligero.
8. Labios brillantes, hidratados por mucho tiempo.
9. Lo que tu madre nunca te contó sobre la higiene íntima.
10. Piel hidratada todo el día.
11. Cuarenta y dos mil novecientas pesetas.
12. Pero no se pueden sustraer al perfume.
13. Base de maquillaje de contorno invisible.
14. Si sólo se piensa en... la última seducción.
15. Una explosión de colores, fuente de inspiración infinita.

Tabla D.2: Valores promedio de RMSE y de ρ para los tres parámetros prosódicos de las frases que forman la prueba subjetiva en el estilo sensual.

	F0 (Hz)		Duración (ms)		Energía (rms)	
Frase	RMSE	ρ	RMSE	ρ	RMSE	ρ
1	26.17	0.19	30.05	0.62	0.030	0.83
2	21.97	0.55	32.01	0.71	0.049	0.75
3	22.79	0.85	26.38	0.55	0.032	0.59
4	19.70	0.58	28.07	0.72	0.052	0.59
5	19.52	0.49	21.94	0.67	0.029	0.78
6	22.55	0.41	21.56	0.71	0.033	0.59
7	27.00	0.32	19.84	0.81	0.025	0.72
8	19.18	0.66	23.42	0.69	0.026	0.60
9	21.51	0.35	25.70	0.70	0.025	0.74
10	27.00	0.50	37.63	0.56	0.014	0.82
11	17.32	0.55	25.81	0.84	0.018	0.77
12	26.85	0.26	25.73	0.64	0.024	0.71
13	26.71	0.33	24.80	0.80	0.016	0.84
14	23.44	0.74	26.79	0.73	0.025	0.55
15	18.55	0.66	29.06	0.73	0.011	0.86

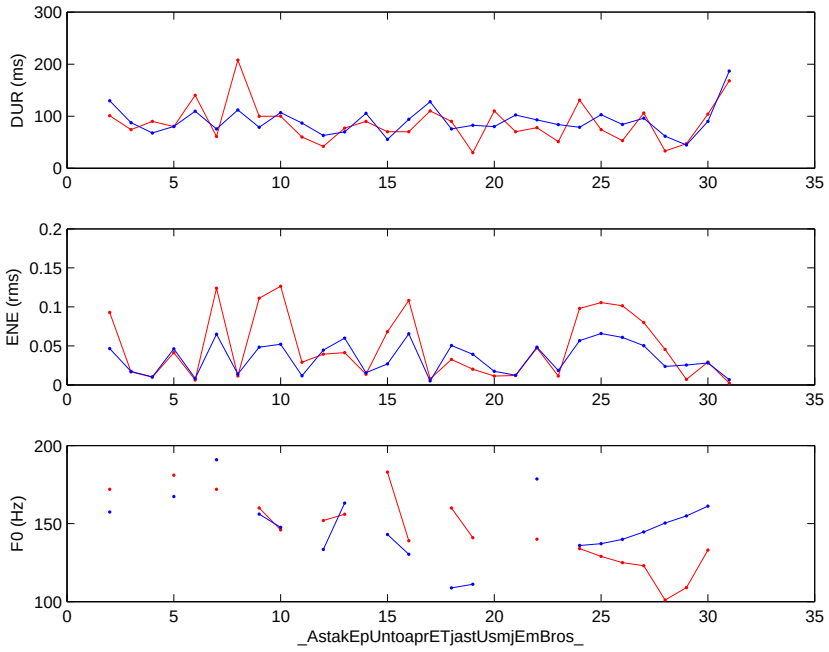


Figura D.16: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 1 del estilo sensual.

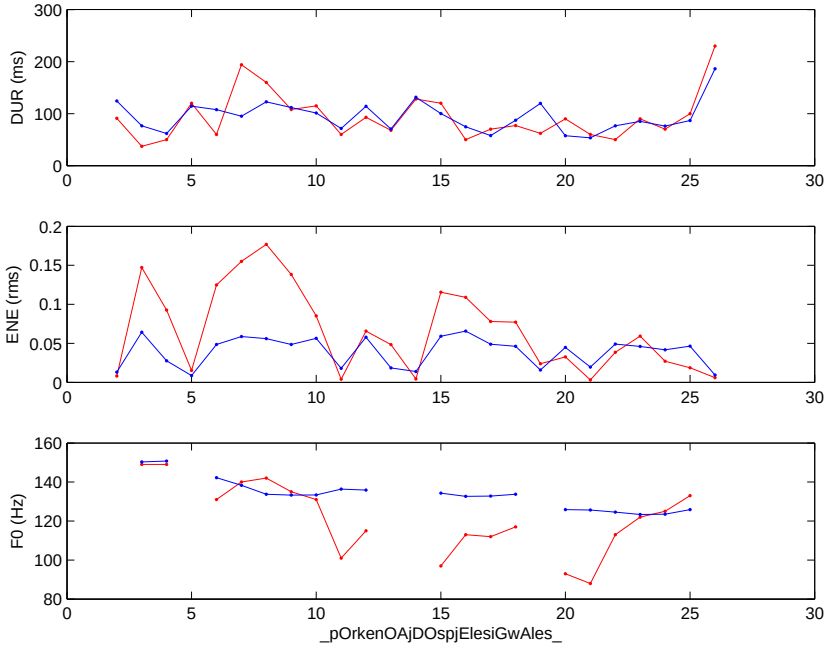


Figura D.17: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 2 del estilo sensual.

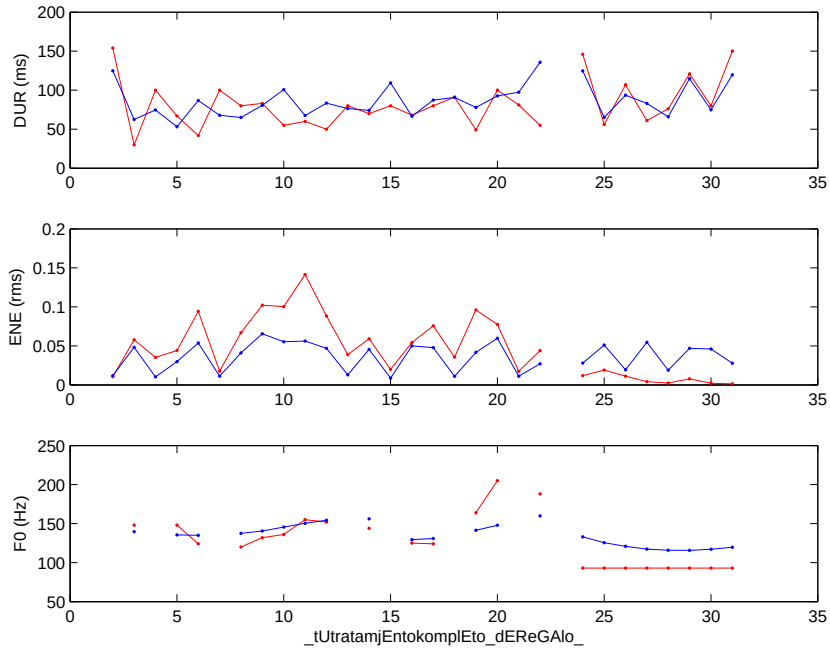


Figura D.18: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 3 del estilo sensual.

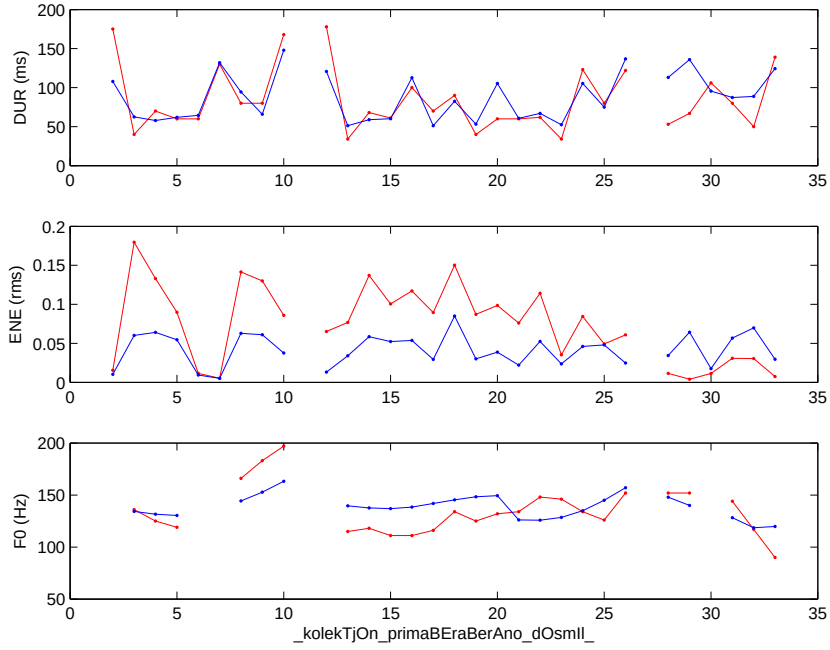


Figura D.19: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 4 del estilo sensual.

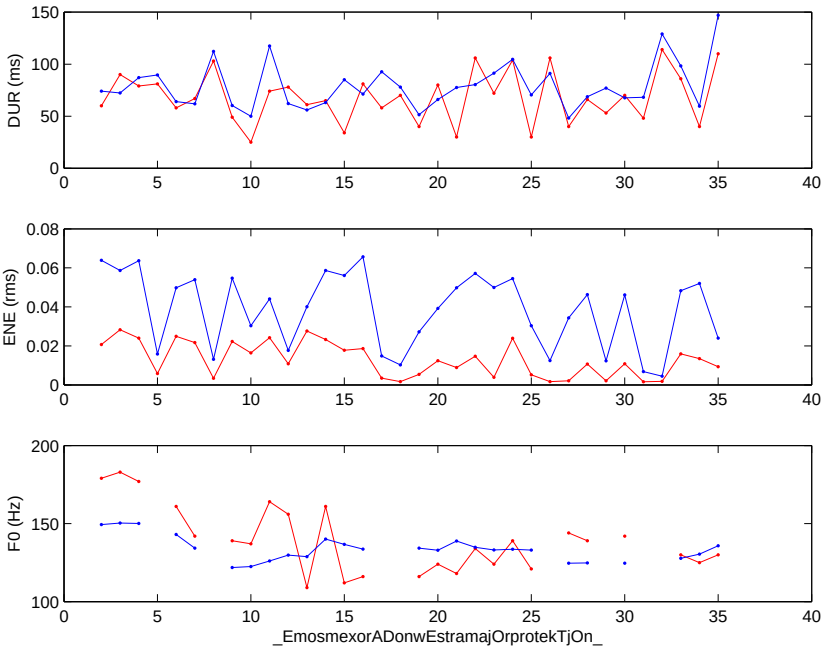


Figura D.20: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 5 del estilo sensual.

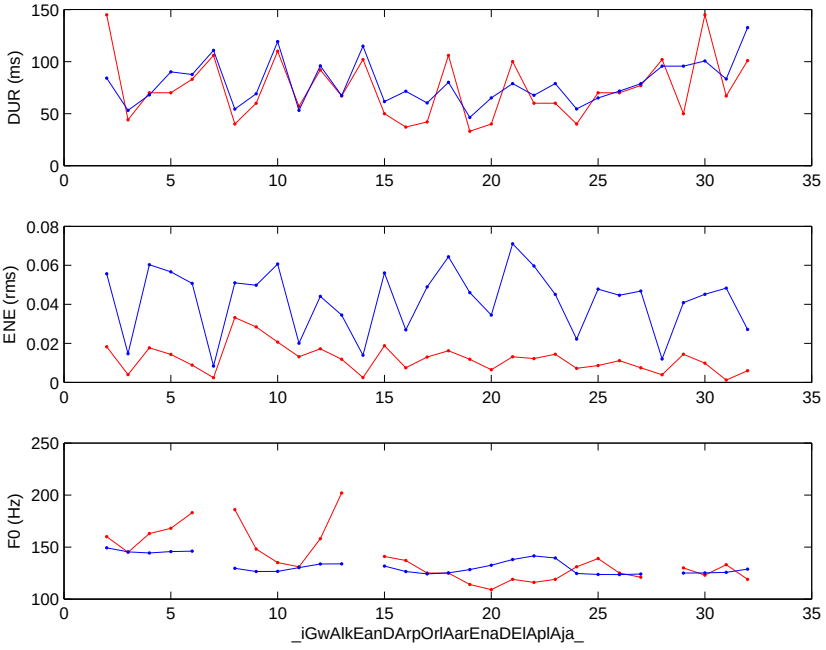


Figura D.21: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 6 del estilo sensual.

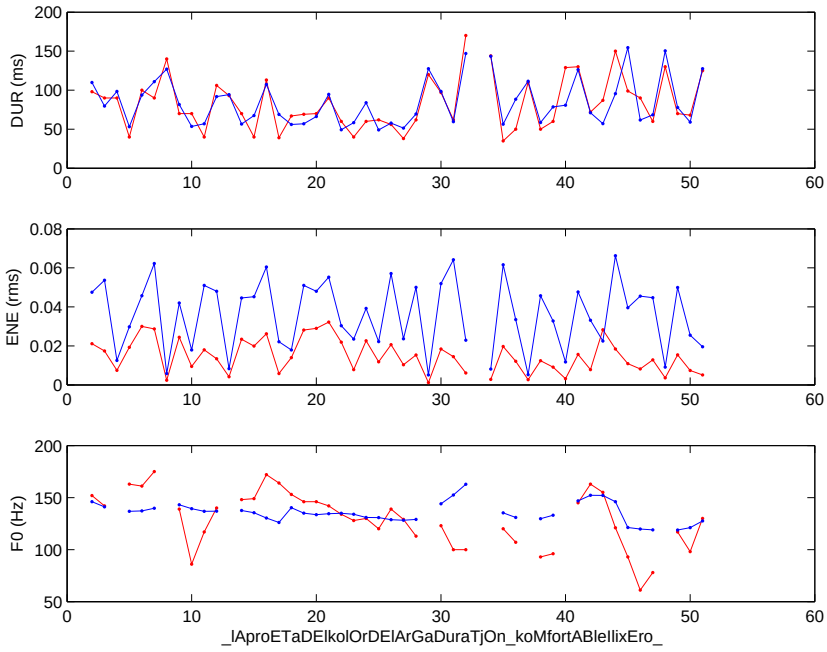


Figura D.22: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 7 del estilo sensual.

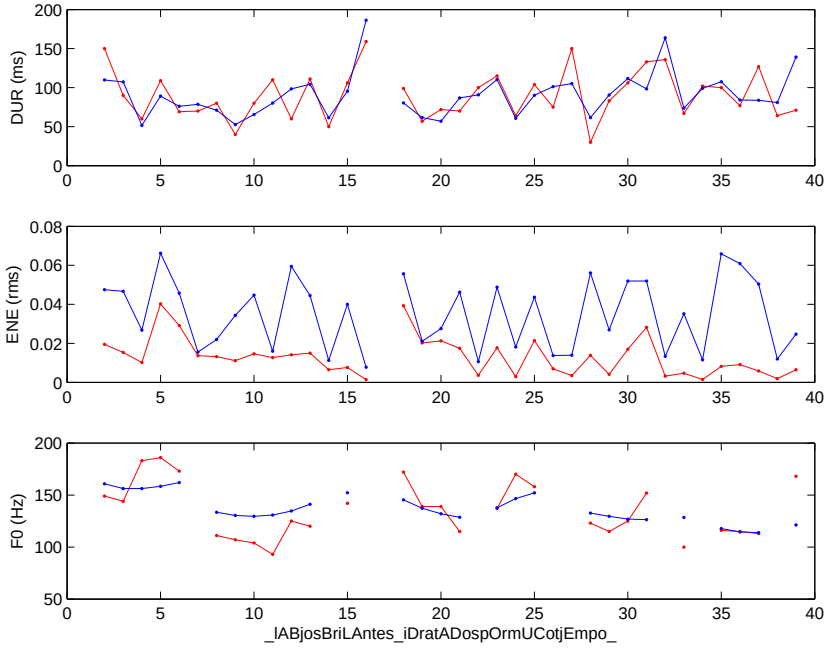


Figura D.23: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 8 del estilo sensual.

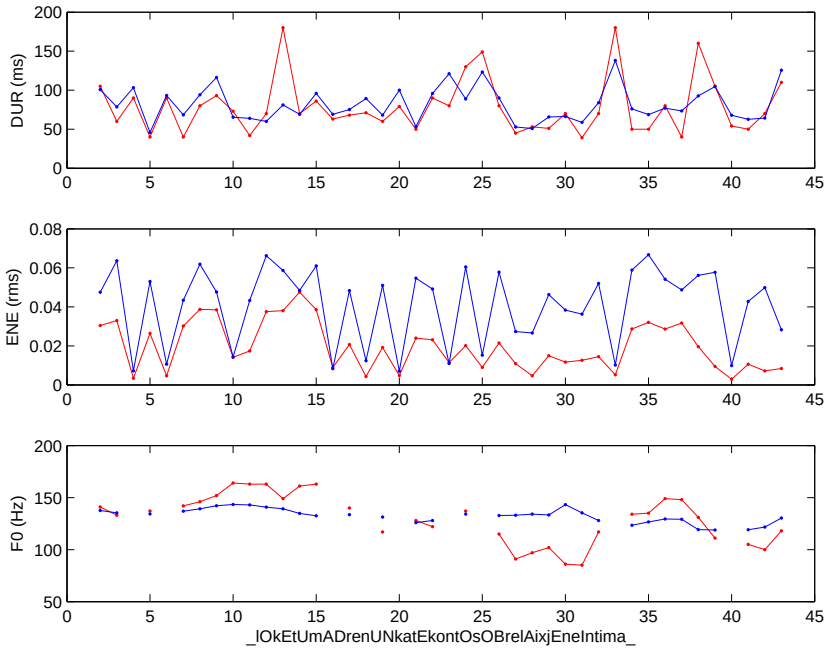


Figura D.24: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 9 del estilo sensual.

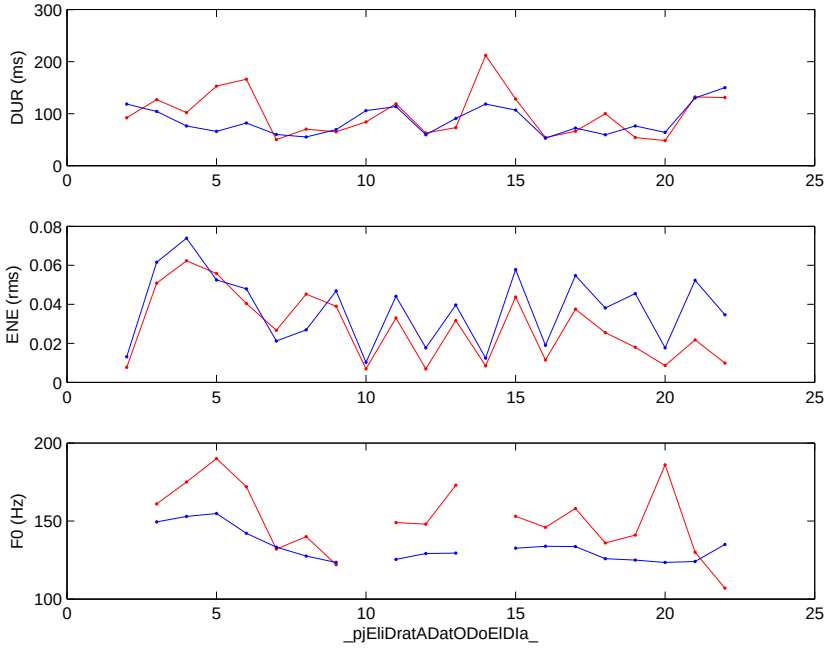


Figura D.25: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 10 del estilo sensual.

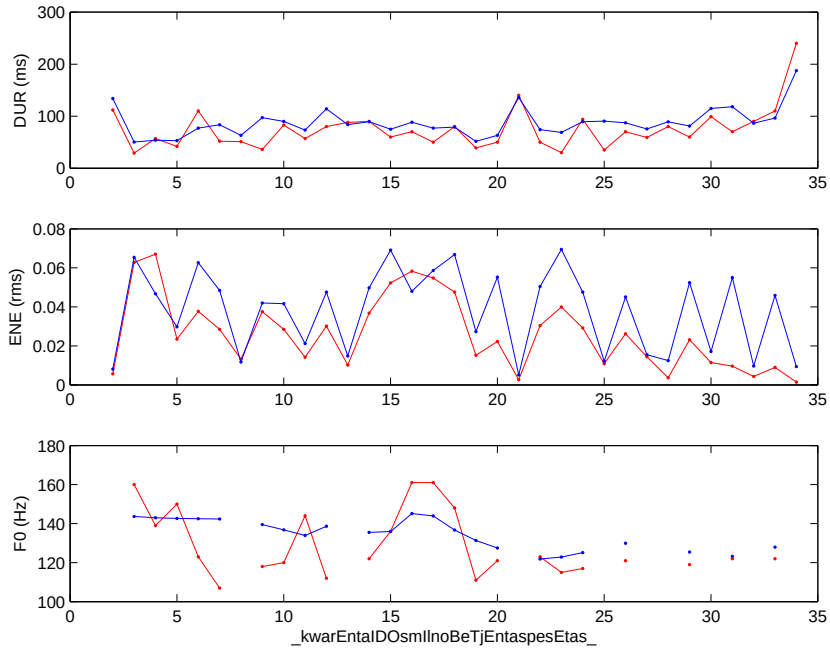


Figura D.26: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 11 del estilo sensual.

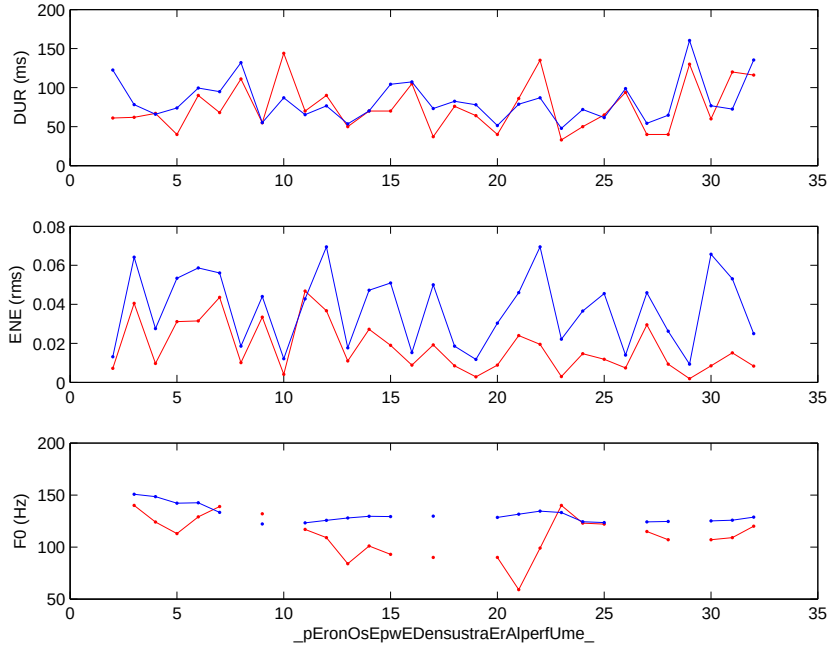


Figura D.27: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 12 del estilo sensual.

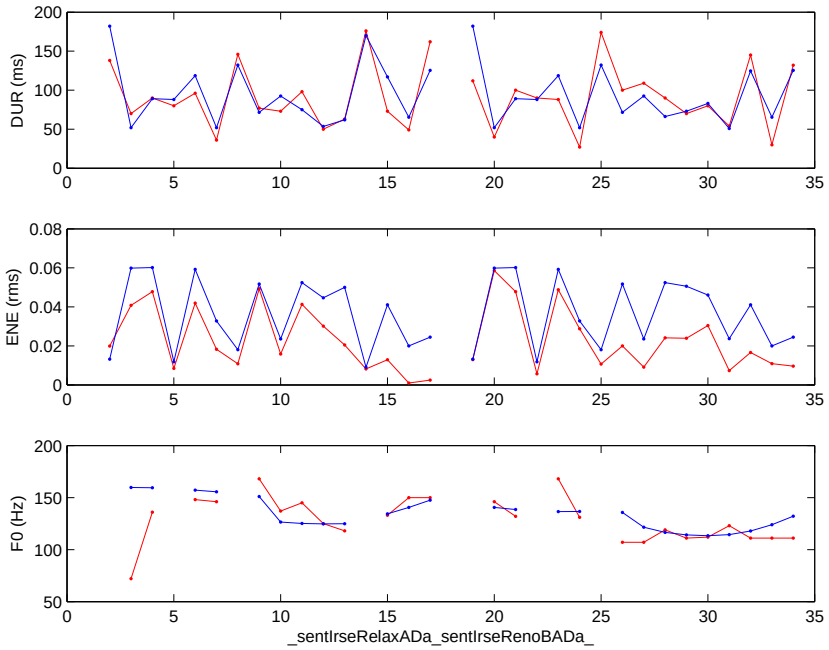


Figura D.28: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 13 del estilo sensual.

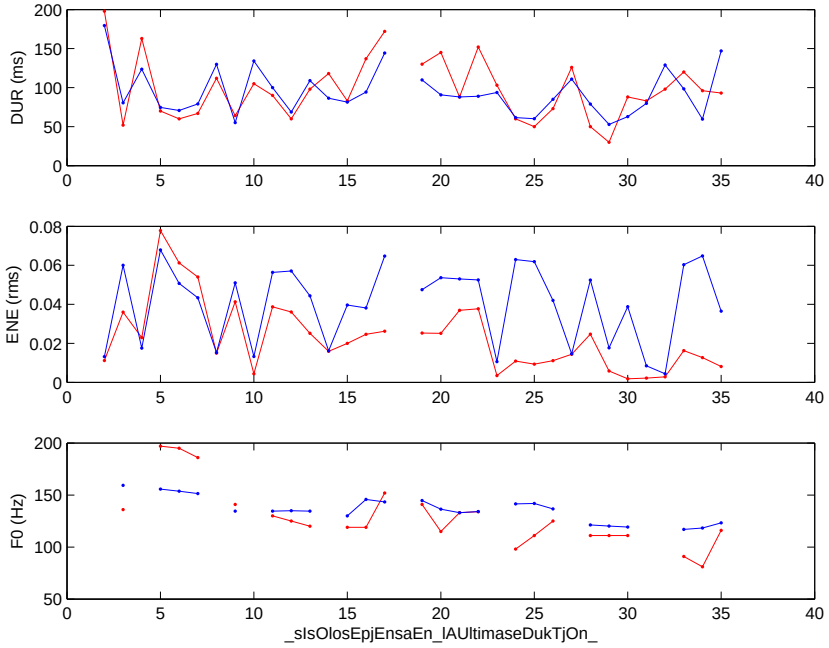


Figura D.29: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 14 del estilo sensual.

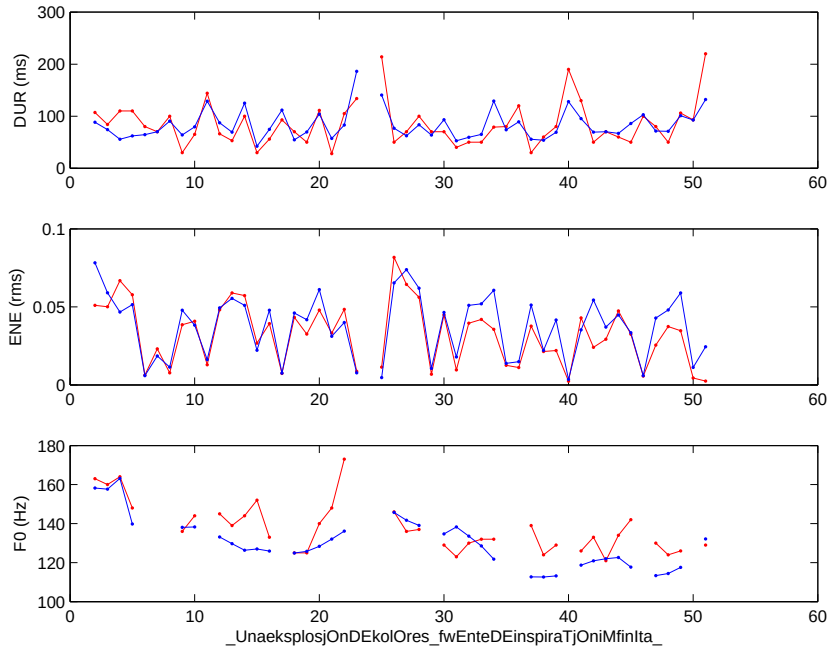


Figura D.30: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 15 del estilo sensual.

D.3. Estilo alegre

Las frases escogidas para la prueba subjetiva de evaluación del módulo de predicción de los parámetros prosódicos en el estilo alegre son las siguientes:

1. La mejor manera de conocer la historia, es divertirse con ella.
2. Quinientos millones por un décimo.
3. Abre a tus hijos, las puertas del mundo.
4. Ahora puedes mejorar tu formación, desde casa.
5. Ahora, te lo ponemos más fácil para aprender.
6. Anúnciese, en el lugar más visitado de Europa.
7. Del presente de sus hijos, depende su futuro.
8. El curso que le abrirá puertas en todo el mundo.
9. Esta es la cola que verás en nuestras pistas.
10. ¡El gran salto para su inglés!
11. He mejorado mi formación desde casa, y a mi ritmo.
12. La educación de su hijo, no debe tener fronteras.
13. Trescientos millones, cambian la vida.
14. La nueva quiniela, se decide en este estadio.
15. Las autoridades advierten, que la lectura de este libro, crea, adicción.

Tabla D.3: Valores promedio de RMSE y de ρ para los tres parámetros prosódicos de las frases que forman la prueba subjetiva en el estilo alegre.

	F0 (Hz)		Duración (ms)		Energía (rms)	
Frase	RMSE	ρ	RMSE	ρ	RMSE	ρ
1	70.43	0.53	32.19	0.67	0.032	0.78
2	82.48	0.03	25.93	0.74	0.030	0.82
3	62.82	0.66	24.46	0.83	0.028	0.72
4	71.39	0.46	20.27	0.85	0.024	0.80
5	62.34	0.83	30.73	0.57	0.021	0.83
6	62.56	0.67	20.87	0.87	0.025	0.70
7	71.35	0.61	19.12	0.84	0.017	0.86
8	74.89	0.73	17.53	0.87	0.023	0.77
9	83.06	0.32	23.57	0.73	0.028	0.68
10	61.64	0.79	29.95	0.79	0.023	0.63
11	83.50	0.25	25.38	0.80	0.020	0.81
12	72.66	0.67	22.81	0.82	0.018	0.87
13	63.61	0.70	16.93	0.87	0.026	0.67
14	85.72	0.72	28.82	0.57	0.028	0.72
15	84.45	0.50	22.24	0.84	0.022	0.76

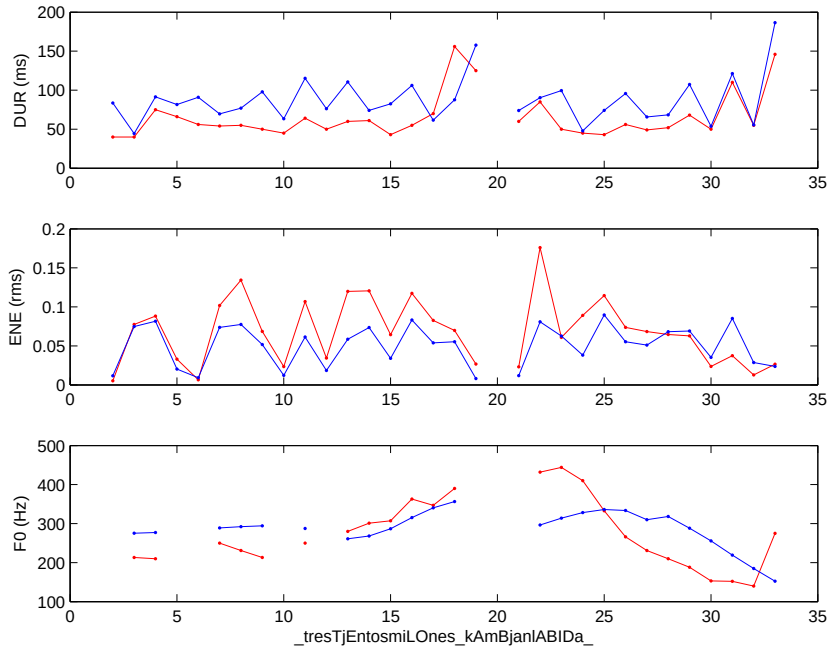


Figura D.31: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 1 del estilo alegre.

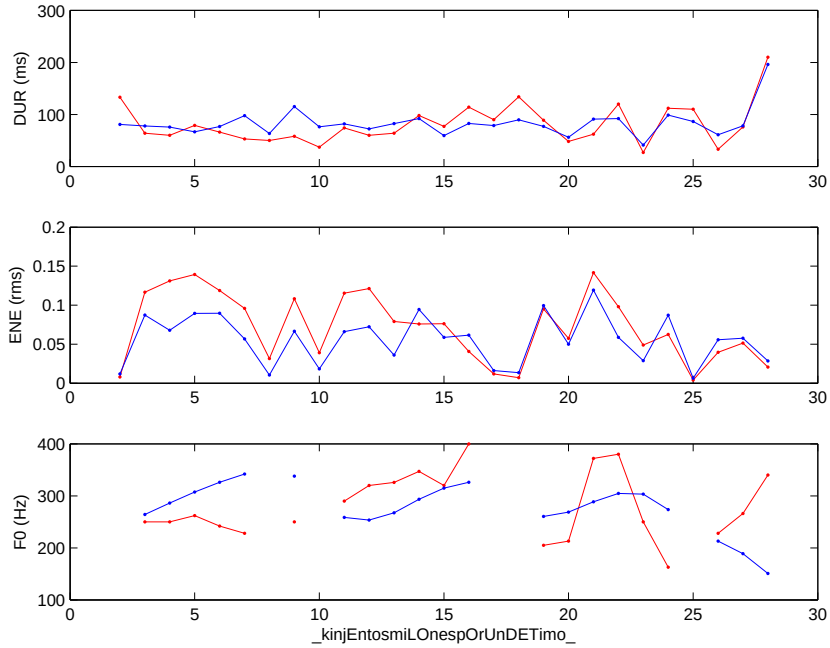


Figura D.32: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 2 del estilo alegre.

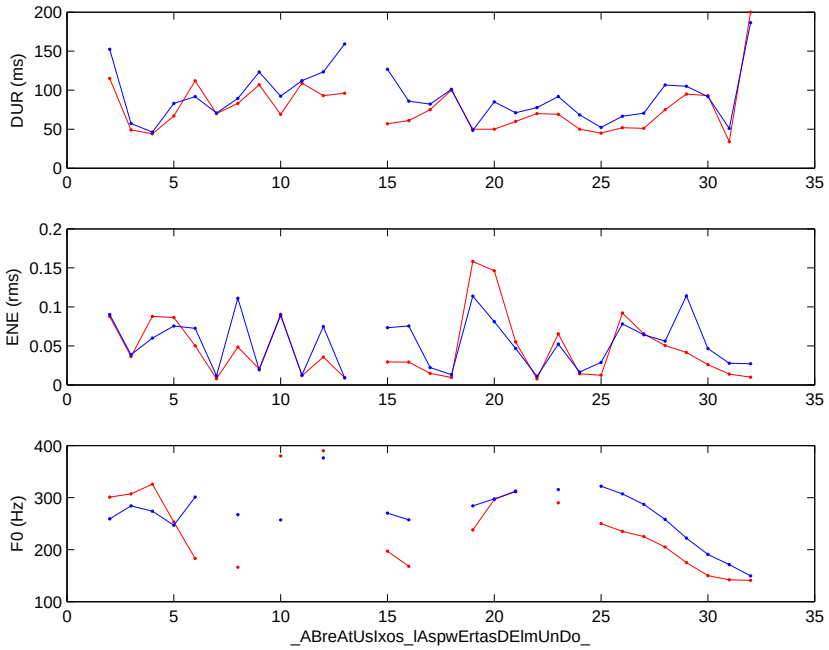


Figura D.33: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 3 del estilo alegre.

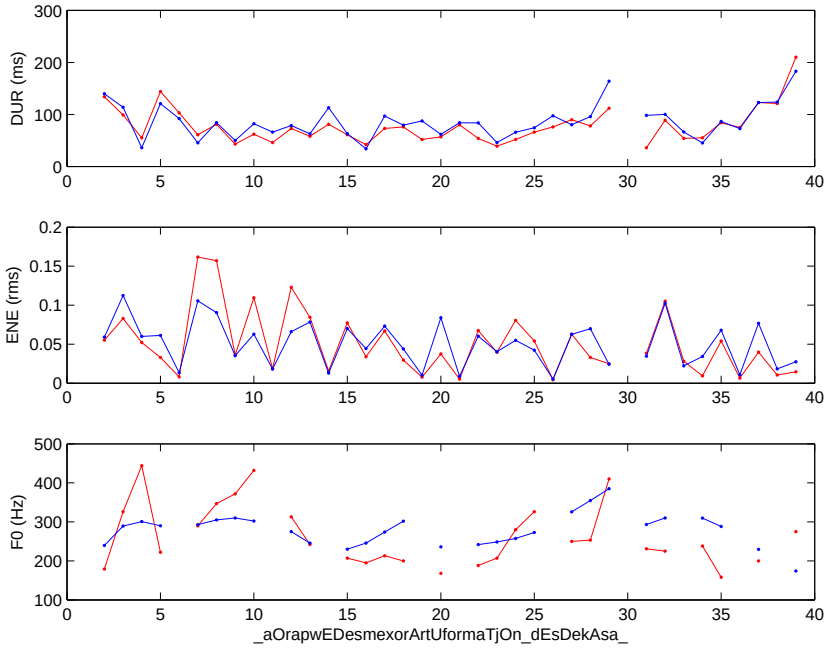


Figura D.34: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 4 del estilo alegre.

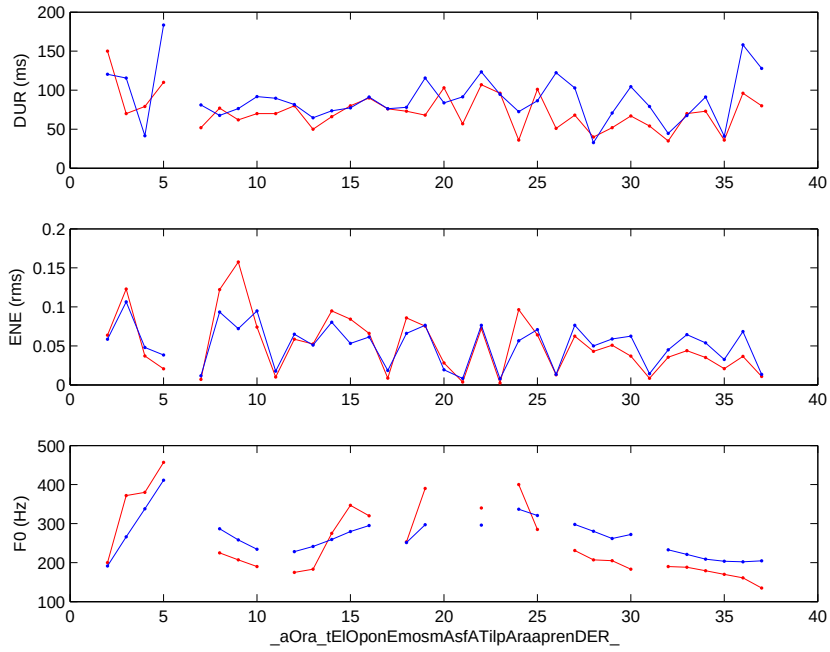


Figura D.35: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 5 del estilo alegre.

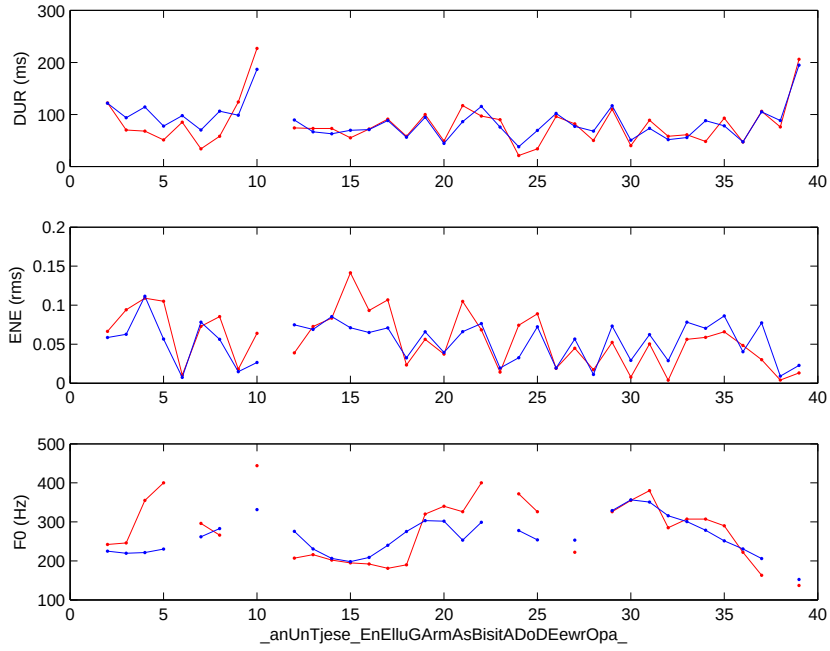


Figura D.36: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 6 del estilo alegre.

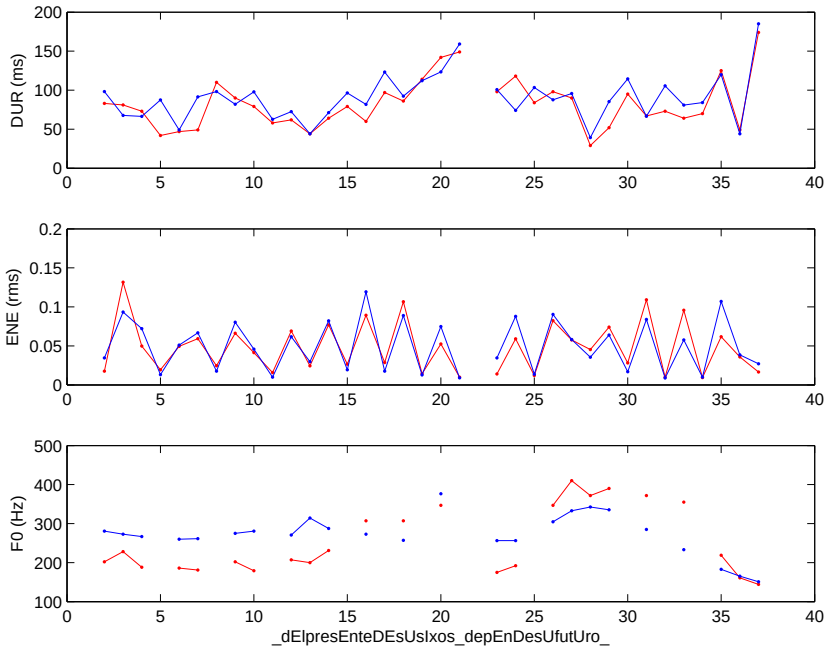


Figura D.37: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 7 del estilo alegre.

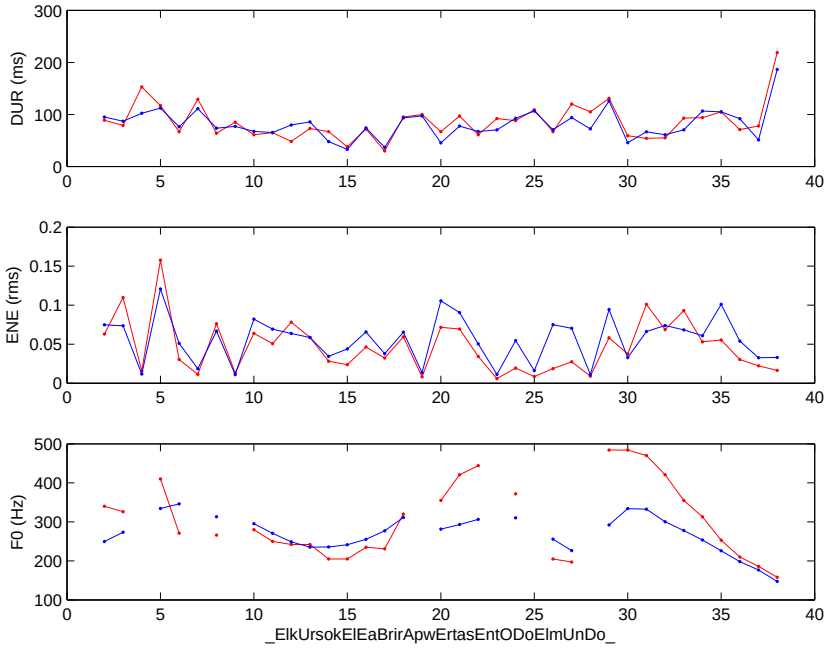


Figura D.38: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 8 del estilo alegre.

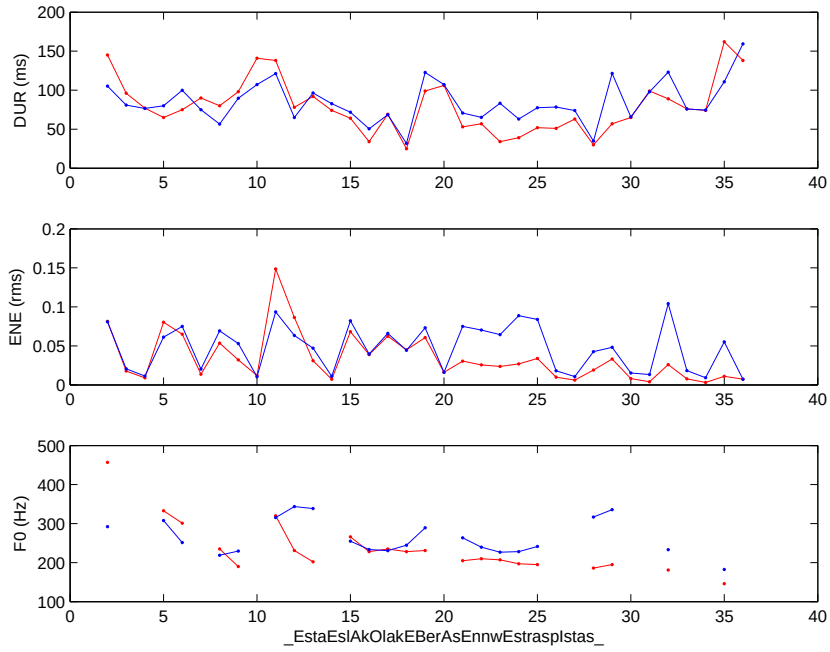


Figura D.39: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 9 del estilo alegre.

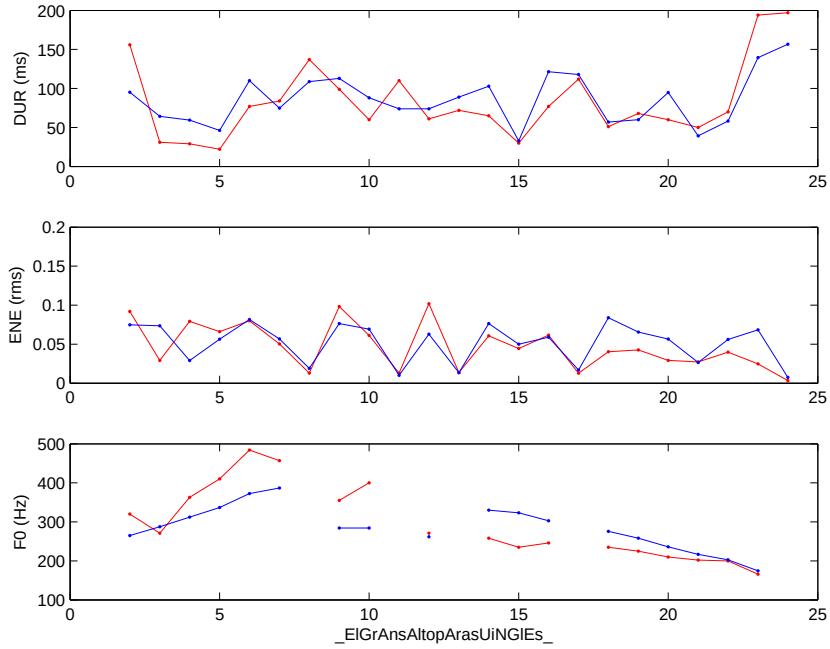


Figura D.40: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 10 del estilo alegre.

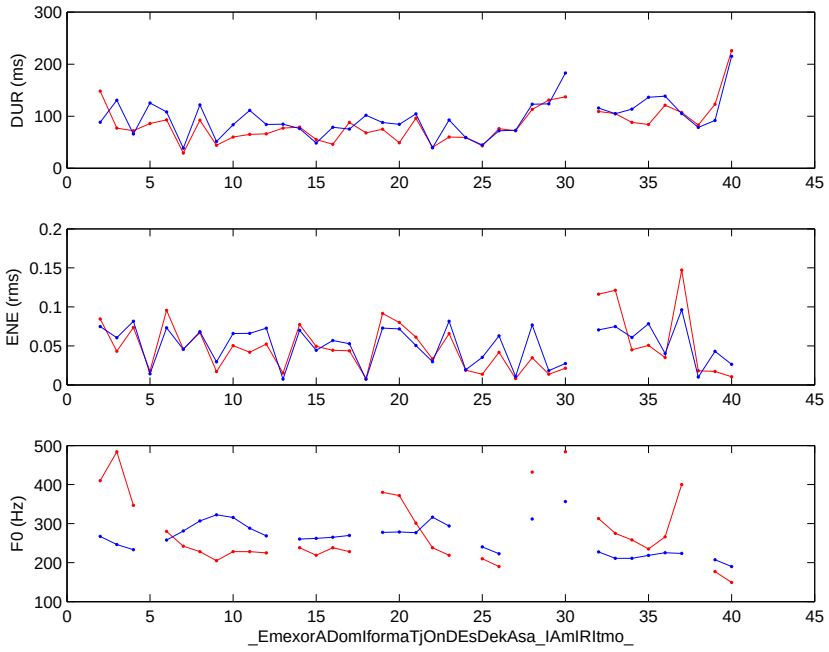


Figura D.41: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 11 del estilo alegre.

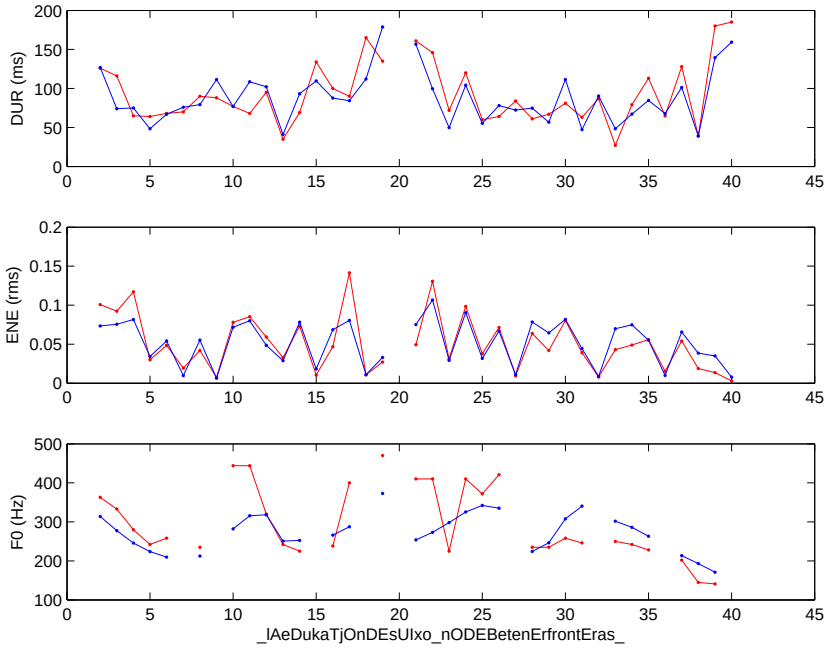


Figura D.42: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 12 del estilo alegre.

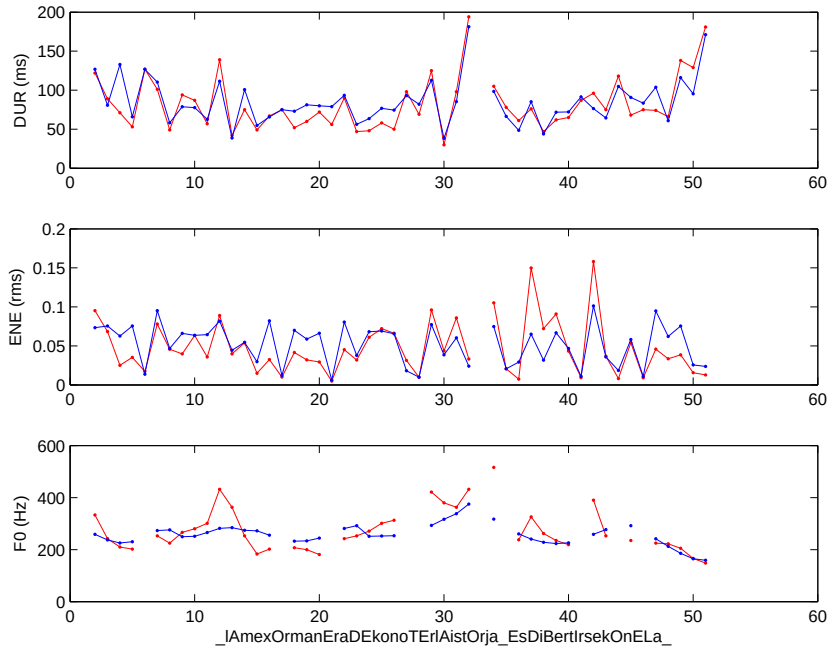


Figura D.43: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 13 del estilo alegre.

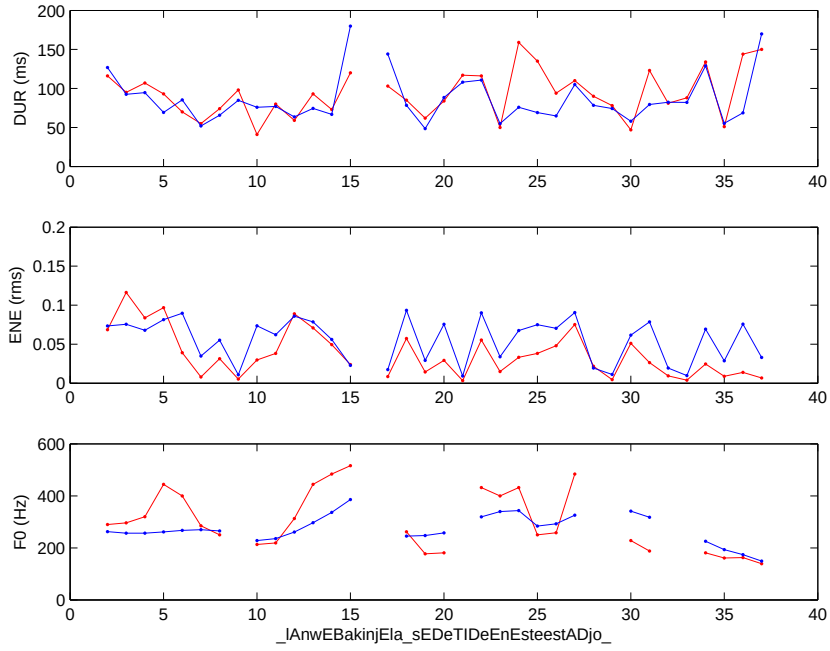


Figura D.44: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 14 del estilo alegre.

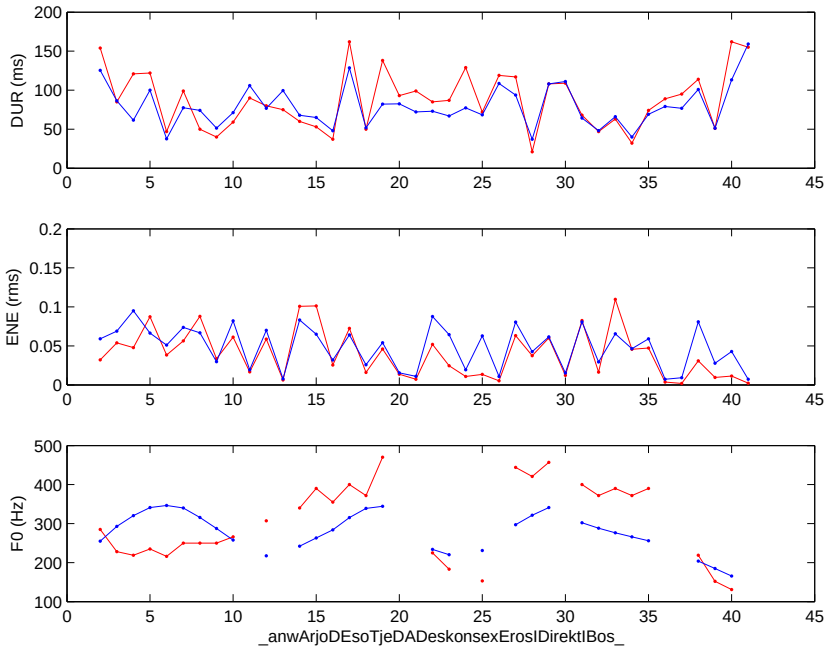


Figura D.45: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 15 del estilo alegre.

D.4. Estilo agresivo

Las frases escogidas para la prueba subjetiva de evaluación del módulo de predicción de los parámetros prosódicos en el estilo agresivo son las siguientes:

1. Crear un automóvil desde cero, está bien.
2. Automóviles que funcionan con hidrógeno, ¿un espejismo? No.
3. ¿De repente tu novio se viene a vivir contigo?
4. Nada debería impedir que cenases siempre con tu familia.
5. Una nueva cumbre en equipamiento, ahora a su alcance.
6. Lo último a caballo entre el campo y la ciudad.
7. Un paisaje que no te esperas. Un perfil deportivo.
8. ¿Qué culpa tienes tú, si conseguiste lo bueno?
9. Rumbo a ti. Se despega de la competencia. Pero nunca del asfalto.
10. Hay quienes saben hacerlo con talento.
11. Lo primero en seguridad. Los Mercedes de nuestro tiempo.
12. Nuevos modelos, nuevos motores y más equipamiento.
13. El secreto de Ferrari. Hay motores, que no envejecen nunca.
14. En este momento están viendo dos coches.
15. Esto es vida. Exclusivo en todos los terrenos.

Tabla D.4: Valores promedio de RMSE y de ρ para los tres parámetros prosódicos de las frases que forman la prueba subjetiva en el estilo agresivo.

	F0 (Hz)		Duración (ms)		Energía (rms)	
Frase	RMSE	ρ	RMSE	ρ	RMSE	ρ
1	51.16	0.51	23.87	0.76	0.024	0.67
2	42.58	0.77	23.13	0.78	0.036	0.72
3	43.75	0.77	30.08	0.87	0.010	0.84
4	58.71	0.73	33.16	0.57	0.013	0.80
5	43.37	0.70	26.36	0.89	0.016	0.81
6	41.70	0.85	29.28	0.59	0.026	0.71
7	49.85	0.81	25.79	0.77	0.016	0.66
8	58.04	0.75	26.70	0.75	0.013	0.79
9	57.77	0.62	42.89	0.55	0.016	0.59
10	57.73	0.68	27.61	0.65	0.010	0.88
11	49.59	0.66	24.45	0.70	0.009	0.84
12	50.63	0.65	34.69	0.59	0.012	0.85
13	42.36	0.72	36.78	0.44	0.011	0.81
14	57.20	0.48	28.24	0.73	0.018	0.71
15	50.81	0.73	34.18	0.44	0.014	0.72

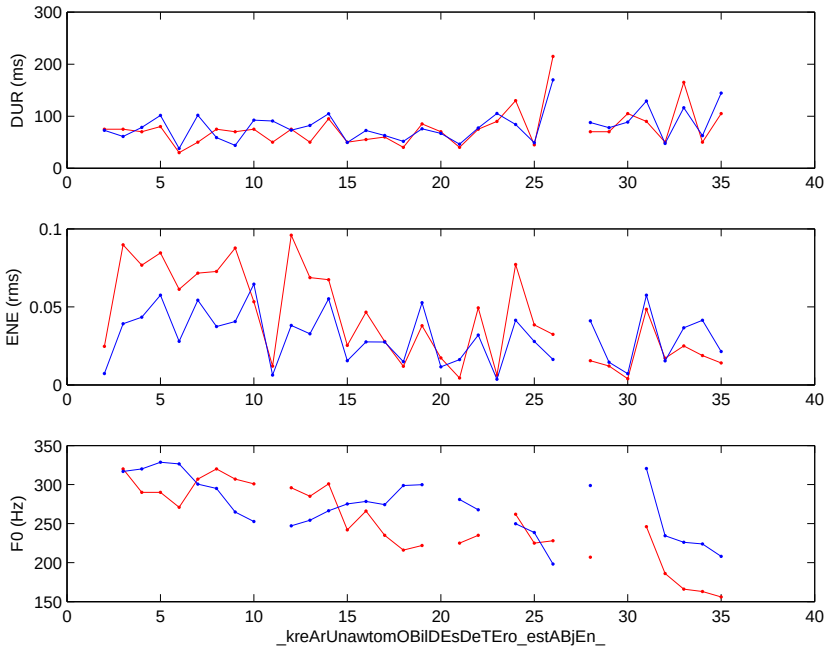


Figura D.46: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 1 del estilo agresivo.

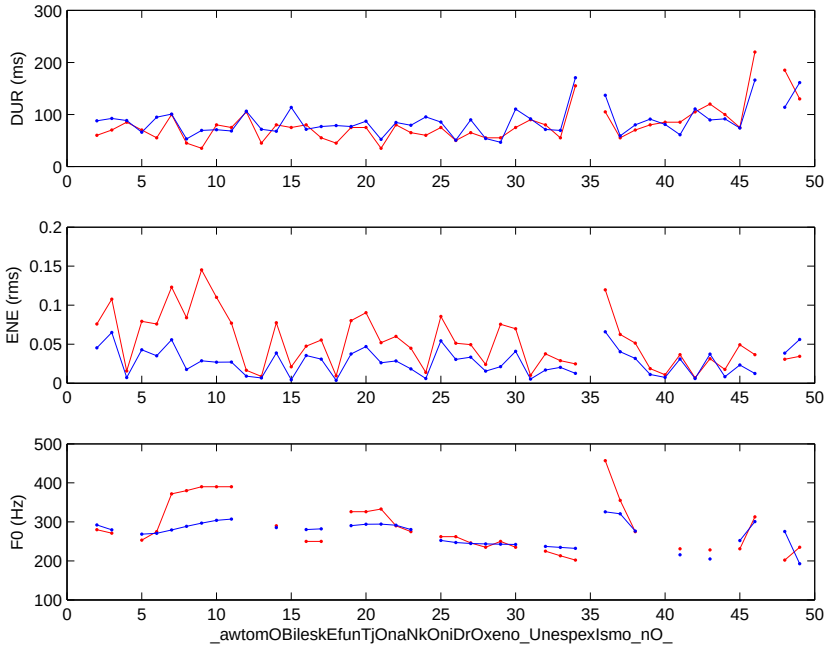


Figura D.47: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 2 del estilo agresivo.

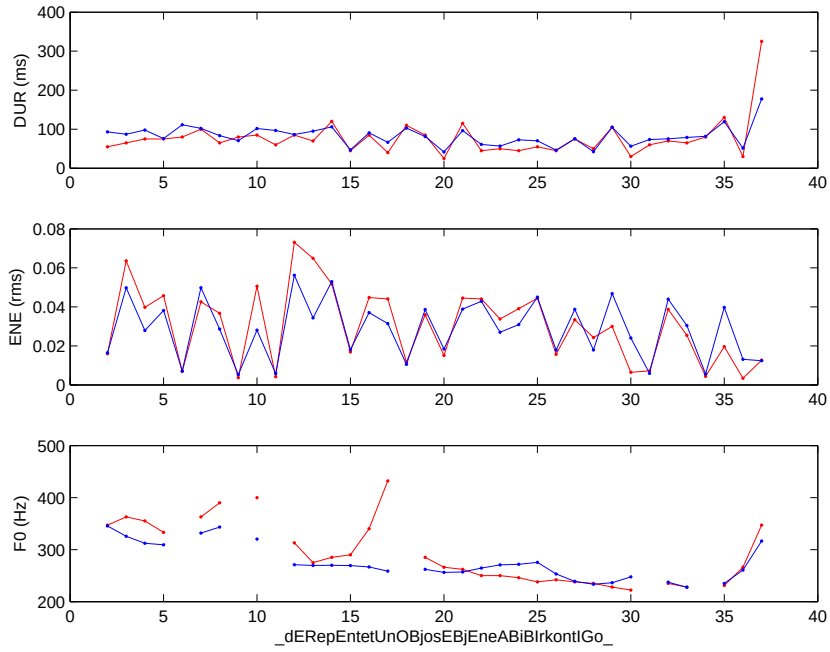


Figura D.48: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 3 del estilo agresivo.

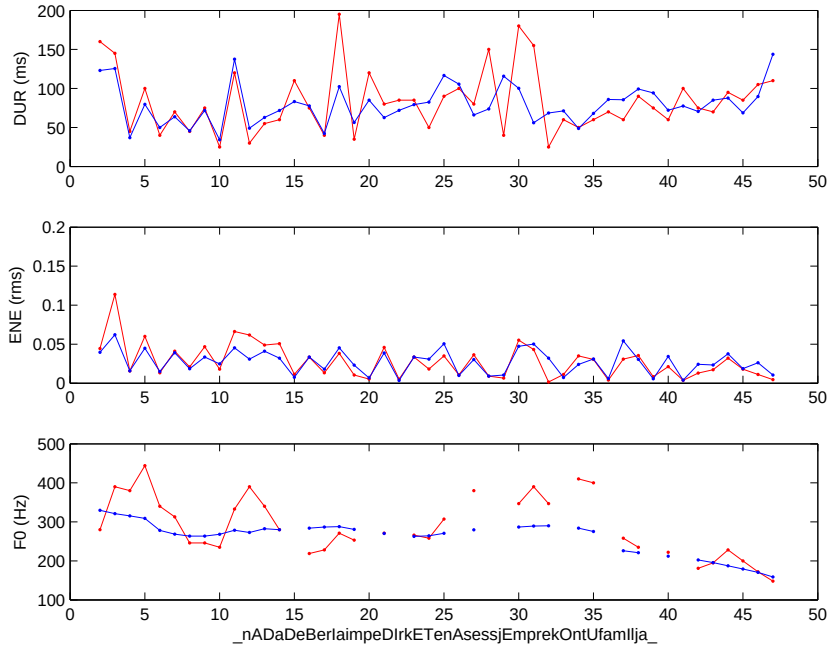


Figura D.49: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 4 del estilo agresivo.

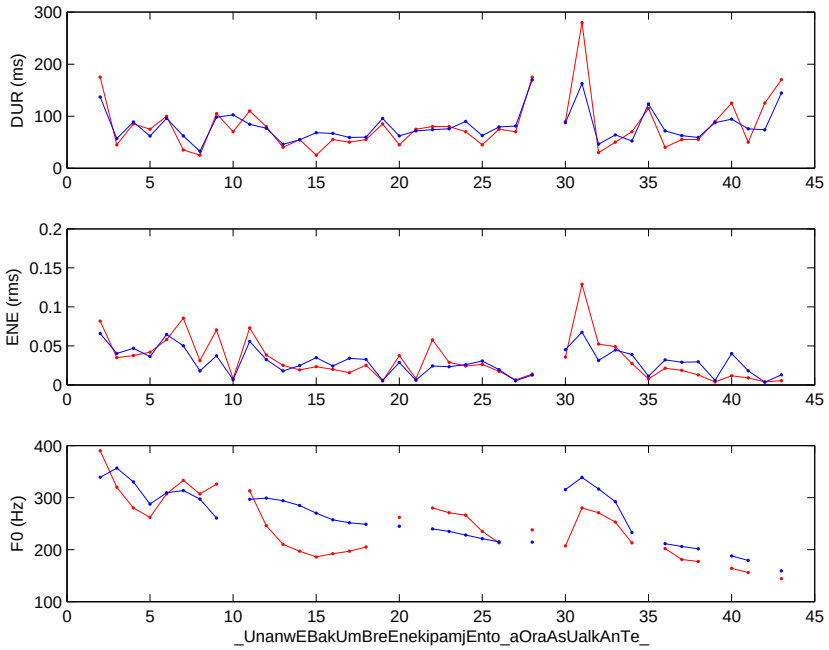


Figura D.50: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 5 del estilo agresivo.

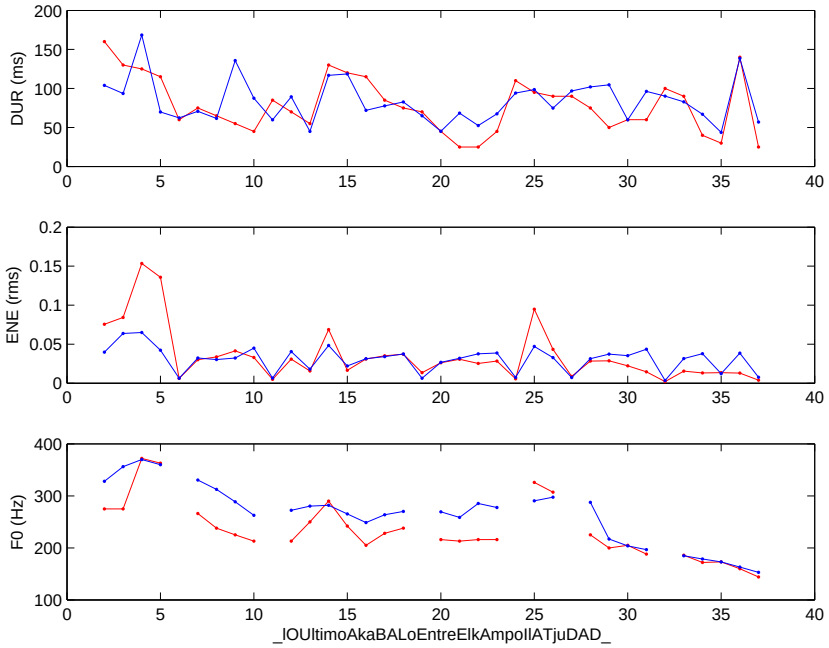


Figura D.51: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 6 del estilo agresivo.

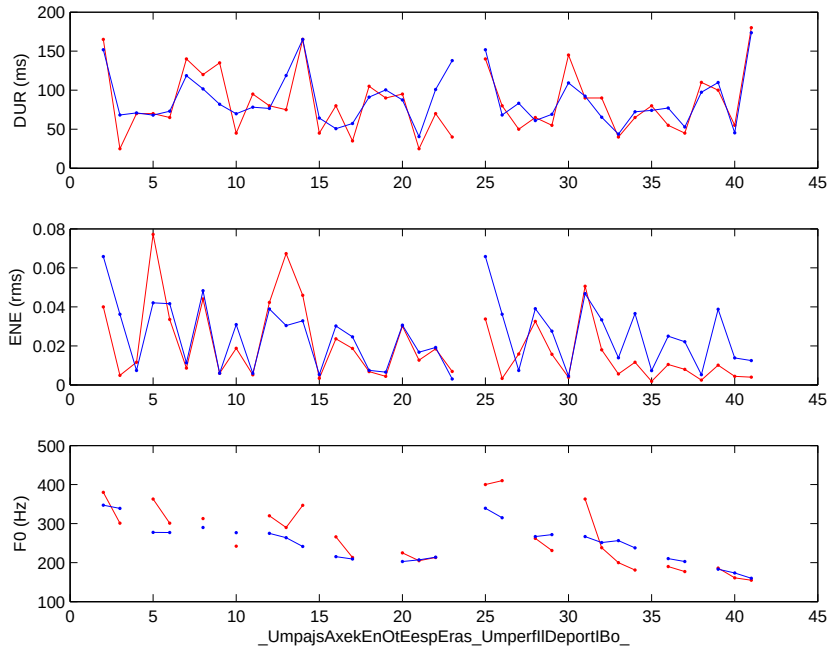


Figura D.52: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 7 del estilo agresivo.

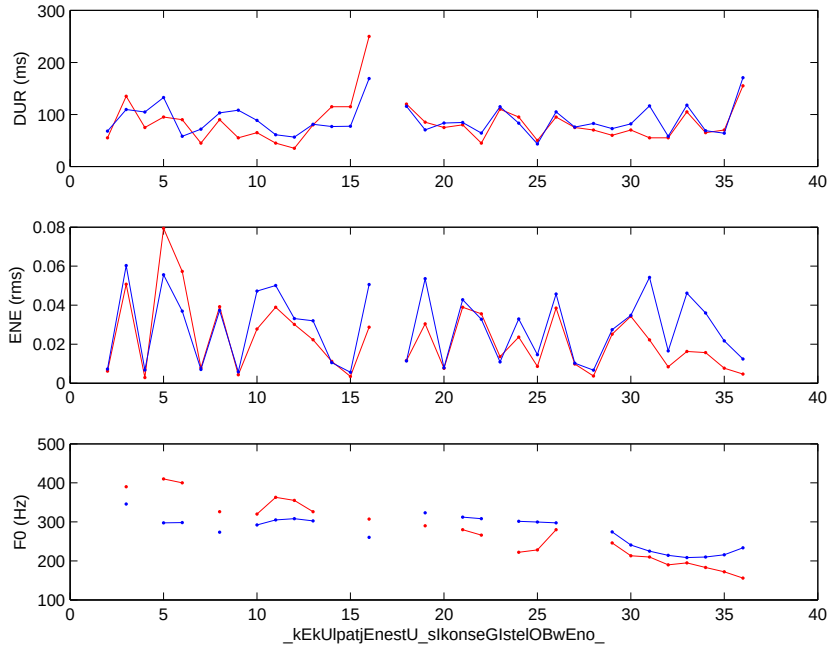


Figura D.53: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 8 del estilo agresivo.

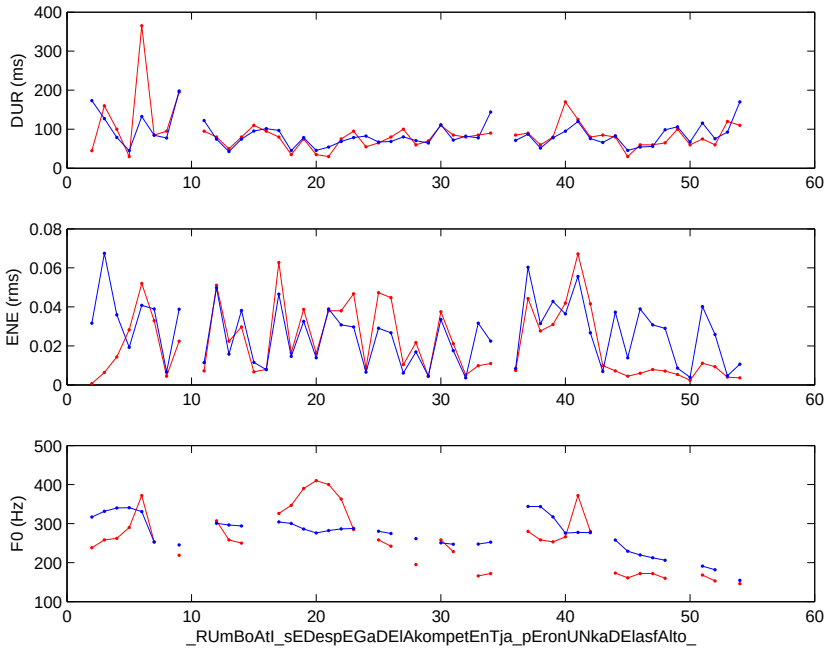


Figura D.54: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 9 del estilo agresivo.

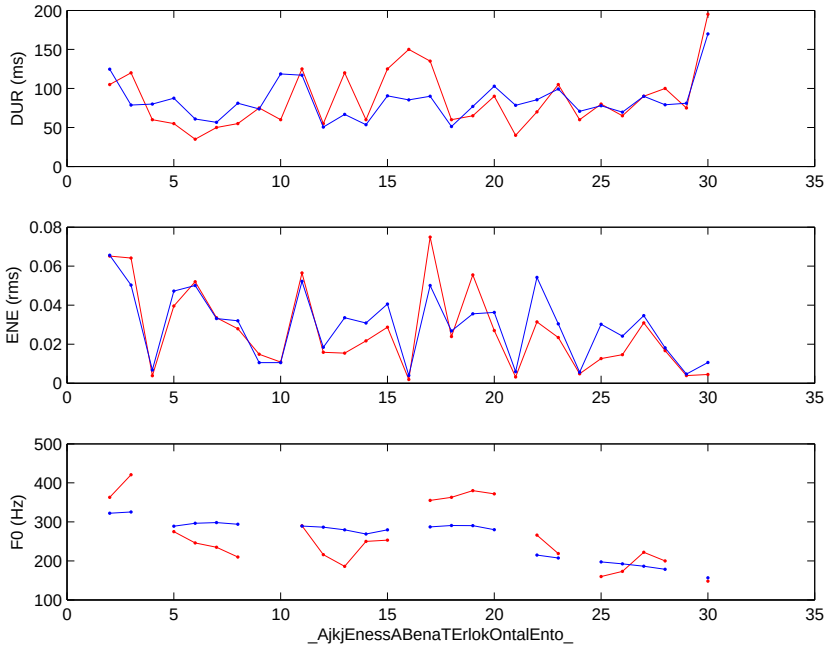


Figura D.55: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 10 del estilo agresivo.

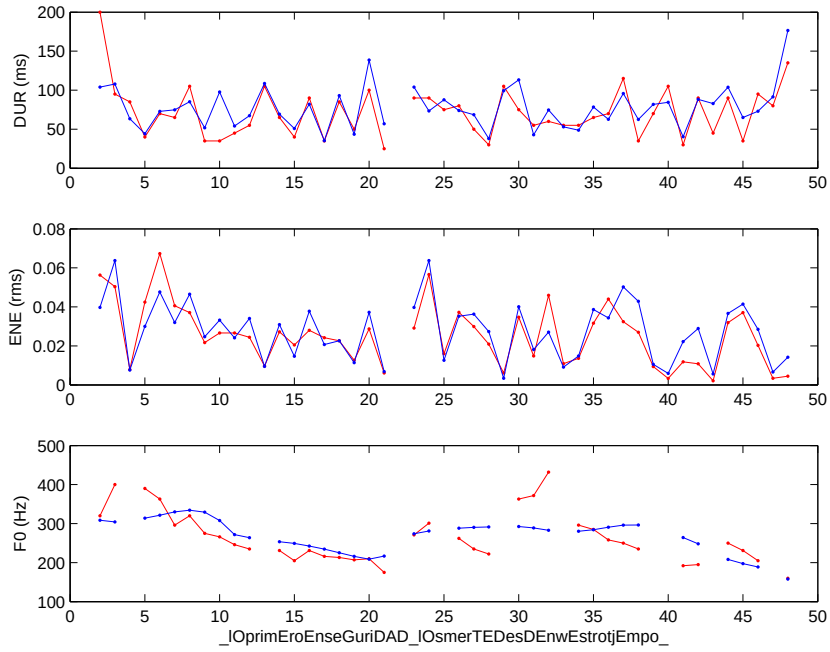


Figura D.56: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 11 del estilo agresivo.

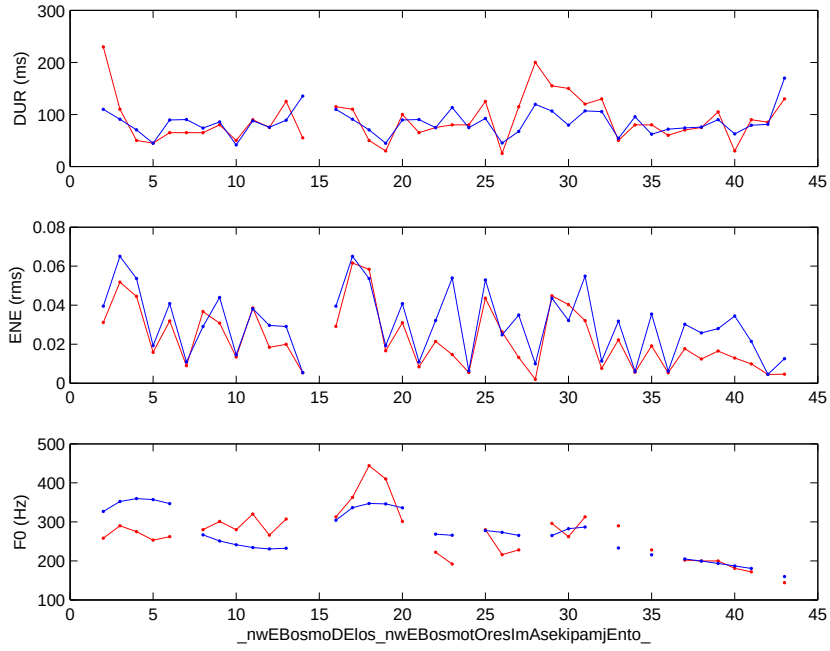


Figura D.57: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 12 del estilo agresivo.

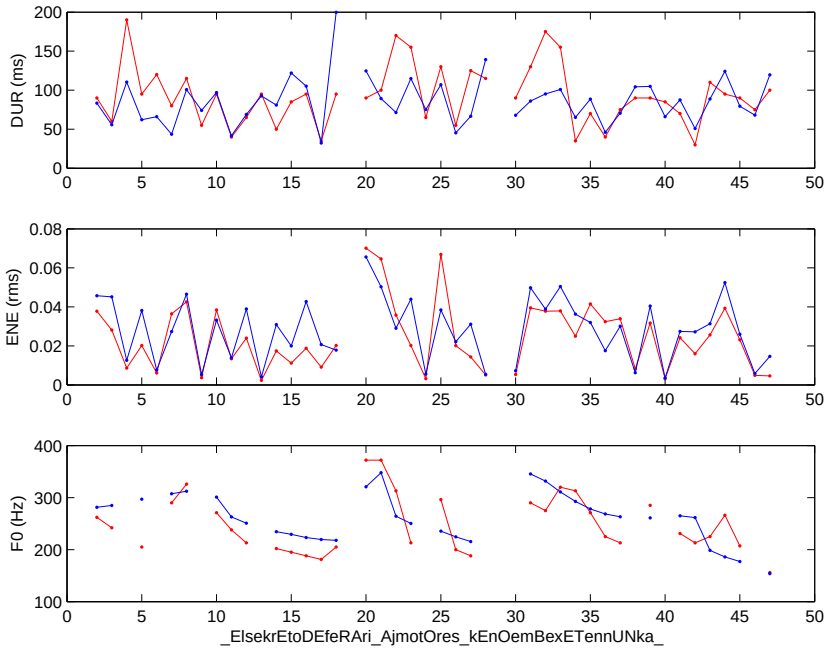


Figura D.58: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 13 del estilo agresivo.

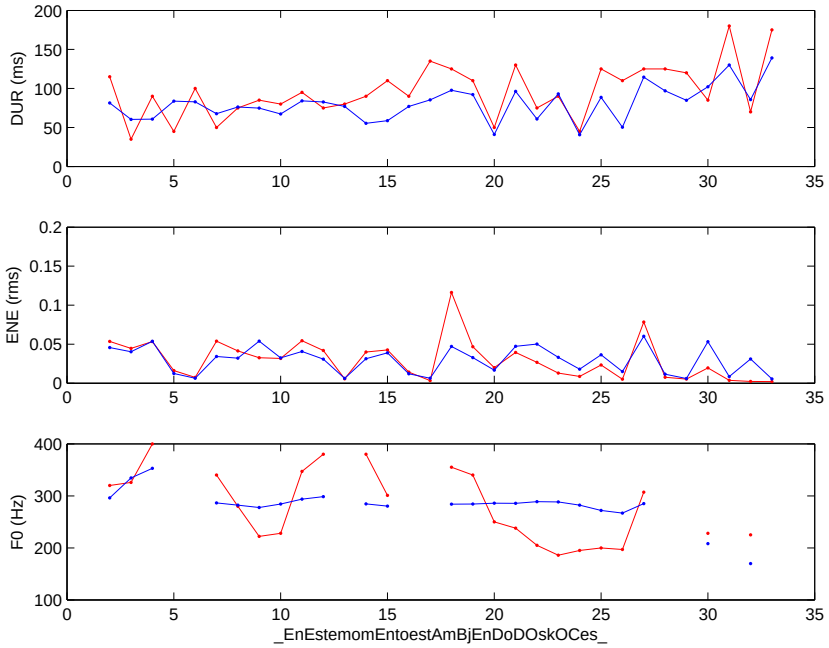


Figura D.59: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 14 del estilo agresivo.

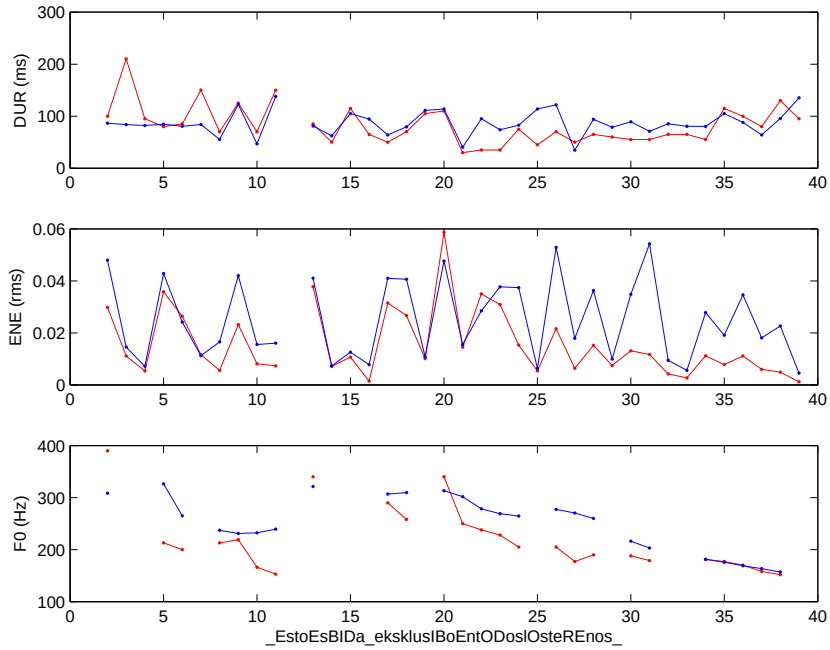


Figura D.60: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 15 del estilo agresivo.

D.5. Estilo triste

Las frases escogidas para la prueba subjetiva de evaluación del módulo de predicción de los parámetros prosódicos en el estilo triste son las siguientes:

1. Líder europeo y grupo mundial de hostelería, y servicios.
2. Nunca el resurgir de un hotel, había sido, tan distinguido.
3. Se permite soñar. Valencia, te sorprenderá.
4. En Cataluña, te sentirás como en casa.
5. Un concepto diferente, en líneas aéreas privadas.
6. Por mar, el viaje es otra cosa.
7. Hospitalidad, desde que calentamos motores.
8. Una forma de trabajar. Un estilo de volar.
9. Día a día, compartiendo ilusiones.
10. La vuelta al mundo en un país. Turquía, naturalmente.
11. Nuestros precios, le quitarán un peso de encima.
12. ¿Soñaba viajes tan especiales a estos precios?
13. Unas vacaciones diferentes.
14. Con nuestras naves descubrirá, un nuevo mundo.
15. Bienvenido, a un mundo con clase.

Tabla D.5: Valores promedio de RMSE y de ρ para los tres parámetros prosódicos de las frases que forman la prueba subjetiva en el estilo triste.

	F0 (Hz)		Duración (ms)		Energía (rms)	
Frase	RMSE	ρ	RMSE	ρ	RMSE	ρ
1	21.54	0.50	51.37	0.48	0.033	0.74
2	16.11	0.51	38.03	0.70	0.030	0.80
3	21.86	0.62	31.99	0.71	0.039	0.77
4	17.60	0.78	55.29	0.30	0.031	0.78
5	21.71	0.76	37.82	0.50	0.039	0.76
6	27.39	0.54	29.88	0.69	0.042	0.78
7	28.33	0.61	36.64	0.27	0.035	0.72
8	17.88	0.66	20.10	0.63	0.031	0.82
9	27.25	0.61	34.41	0.63	0.037	0.70
10	17.19	0.70	36.08	0.61	0.032	0.79
11	26.43	0.34	25.42	0.61	0.049	0.65
12	28.99	0.25	27.12	0.51	0.025	0.88
13	21.72	0.62	21.85	0.70	0.031	0.87
14	16.20	0.53	38.88	0.42	0.035	0.76
15	22.50	0.74	56.05	0.46	0.049	0.74

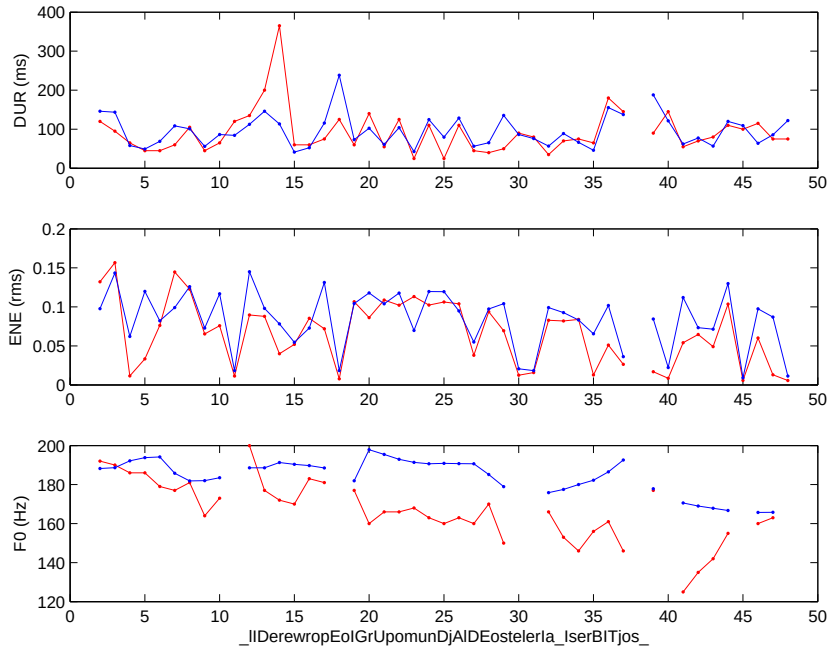


Figura D.61: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 1 del estilo triste.

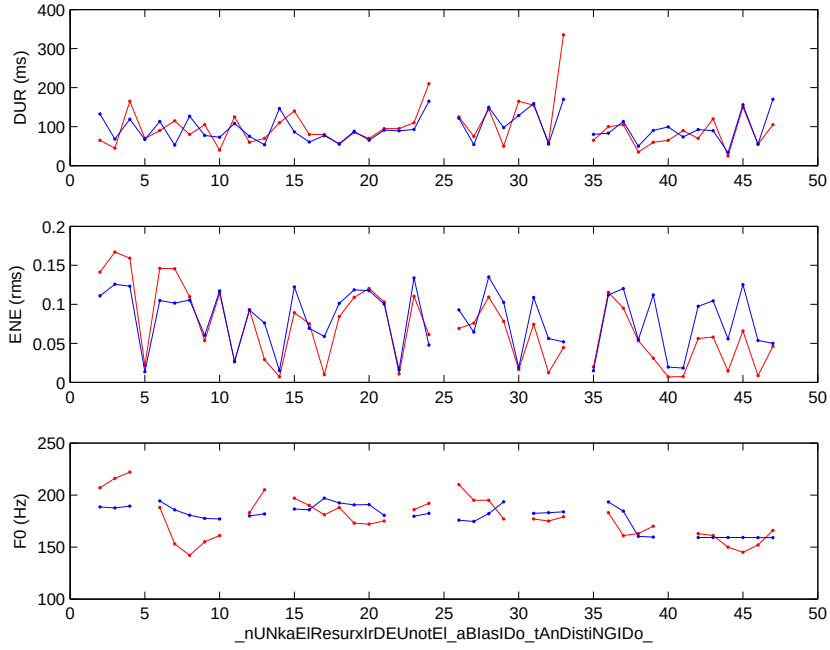


Figura D.62: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 2 del estilo triste.

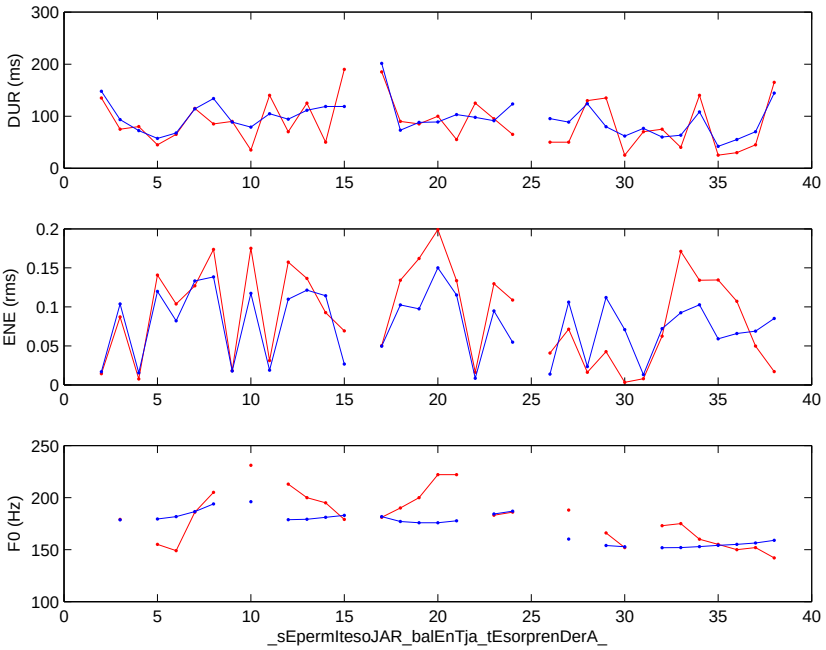


Figura D.63: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 3 del estilo triste.

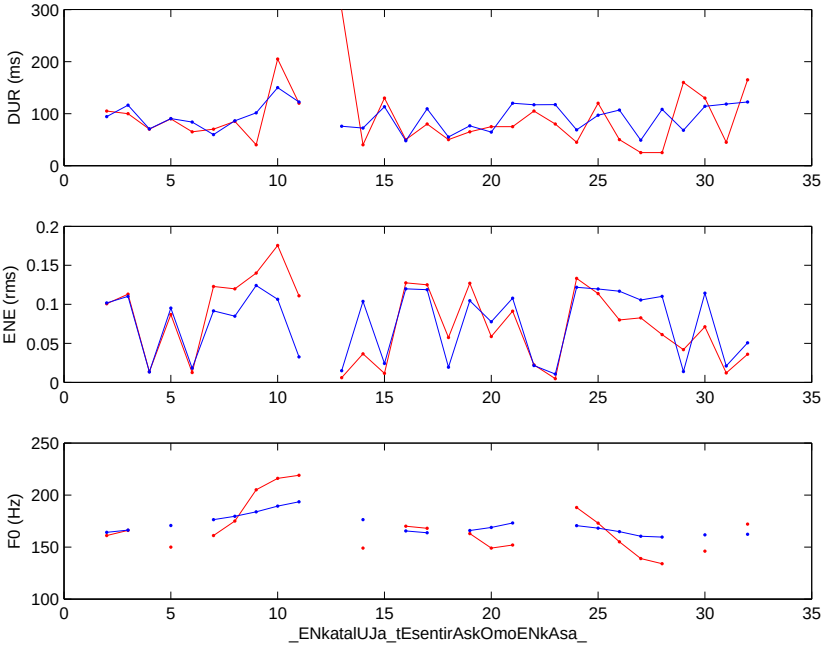


Figura D.64: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 4 del estilo triste.

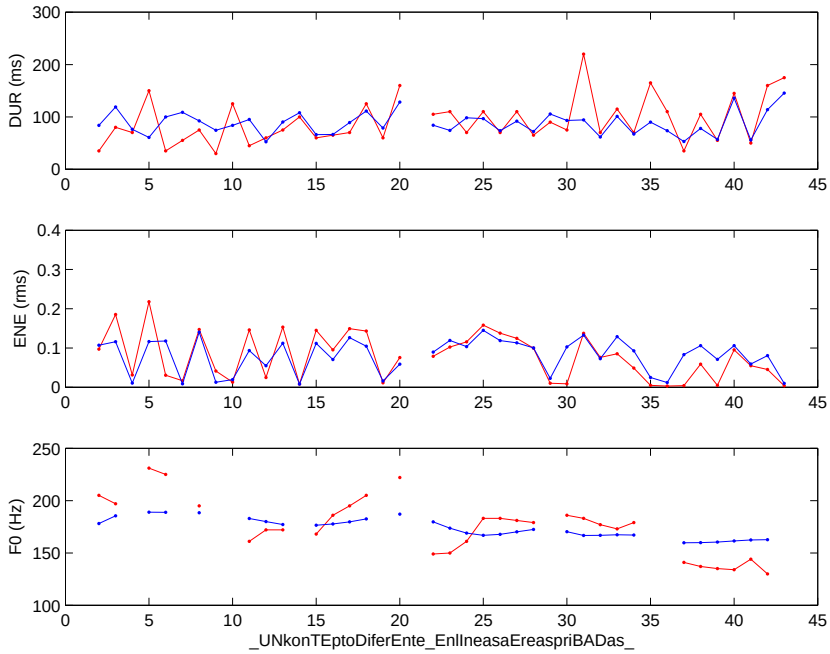


Figura D.65: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 5 del estilo triste.

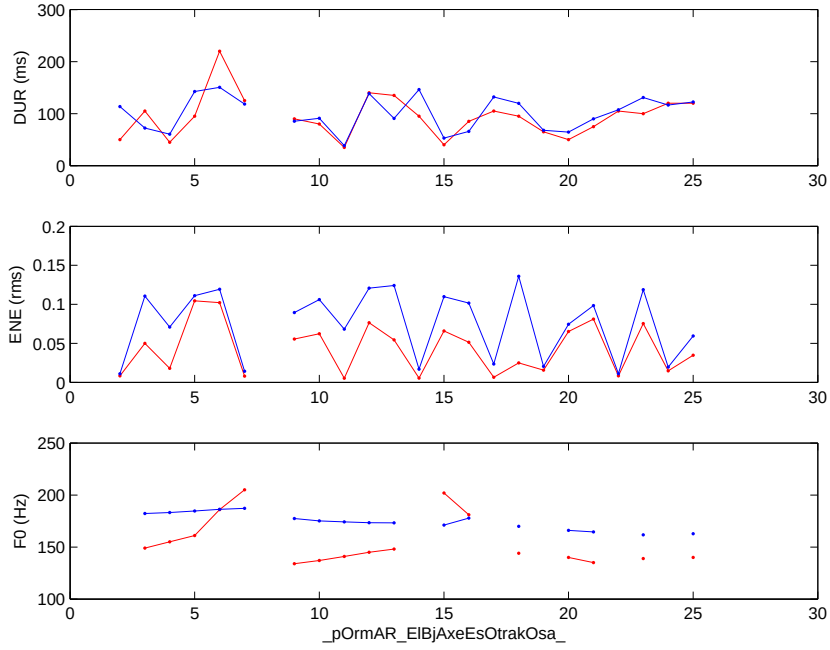


Figura D.66: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 6 del estilo triste.

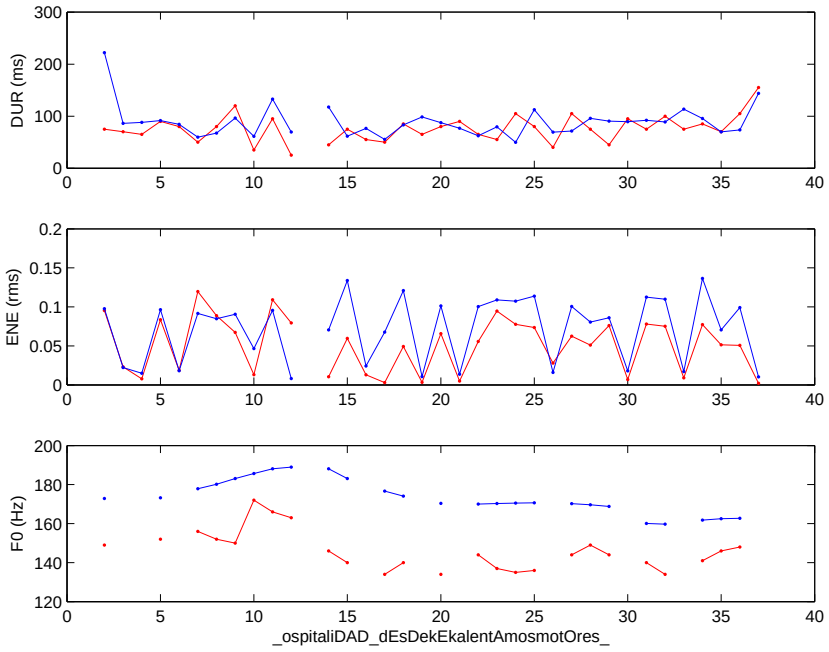


Figura D.67: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 7 del estilo triste.

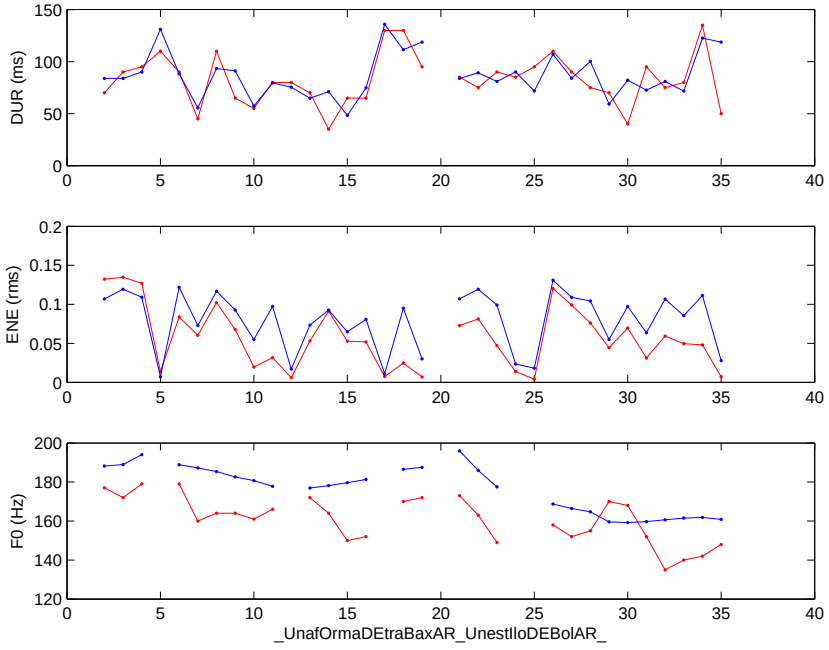


Figura D.68: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 8 del estilo triste.

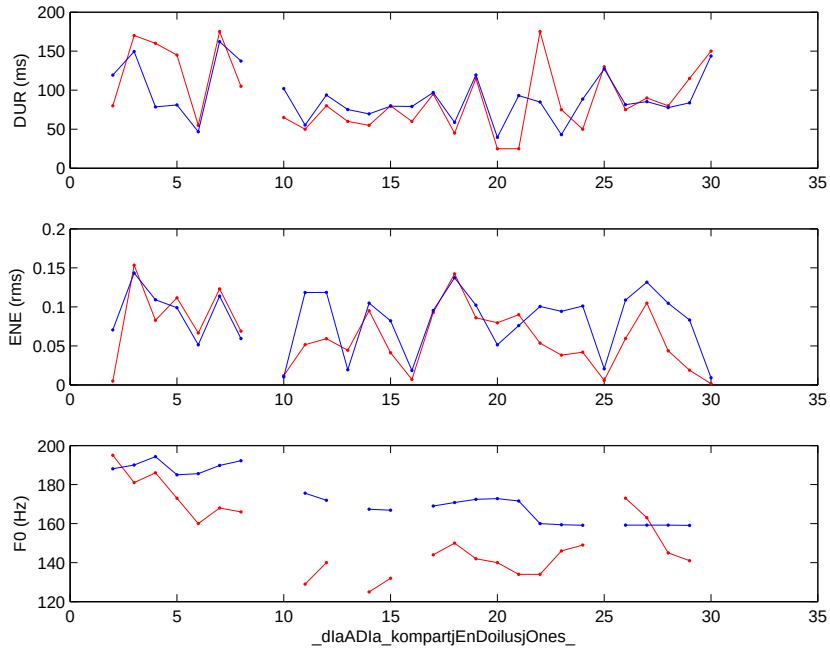


Figura D.69: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 9 del estilo triste.

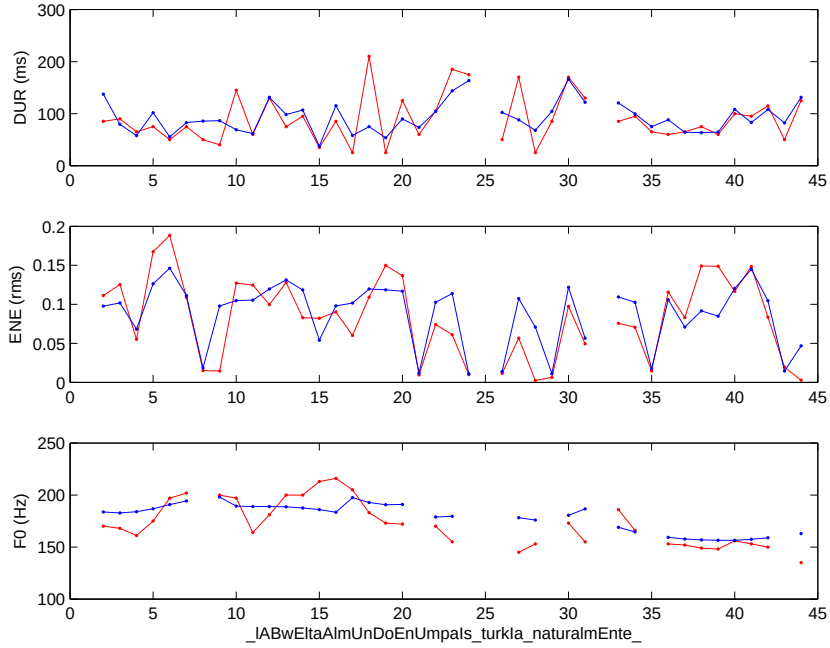


Figura D.70: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 10 del estilo triste.

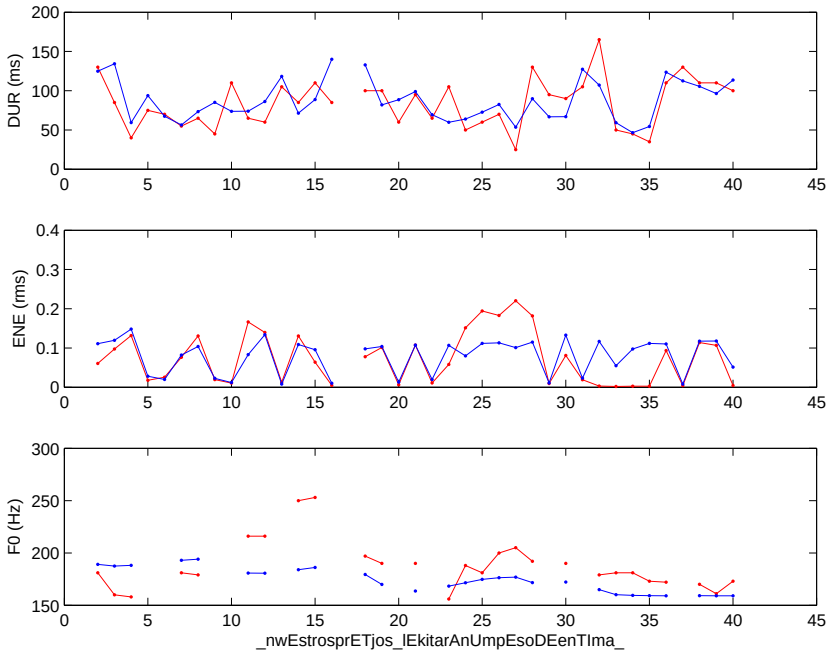


Figura D.71: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 11 del estilo triste.

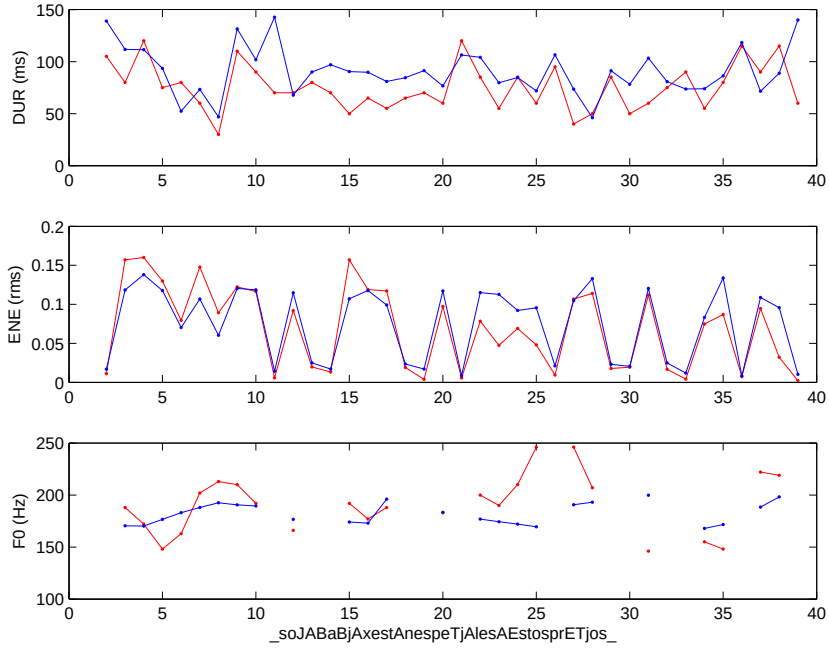


Figura D.72: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 12 del estilo triste.

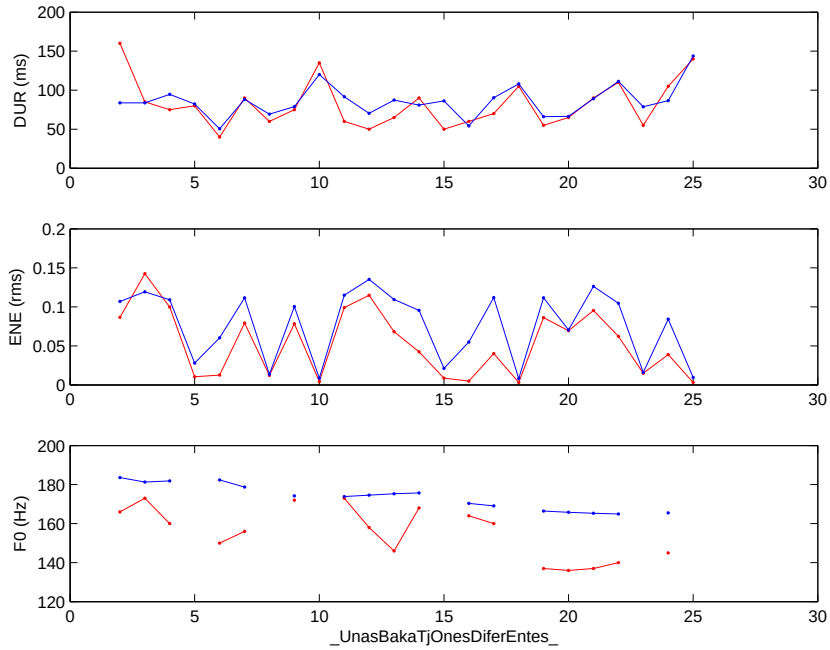


Figura D.73: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 13 del estilo triste.

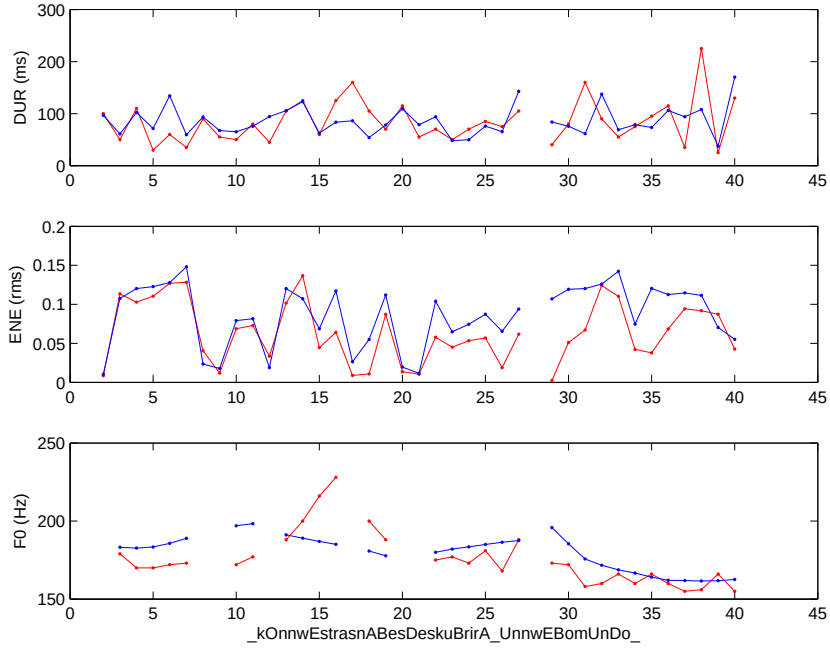


Figura D.74: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 14 del estilo triste.

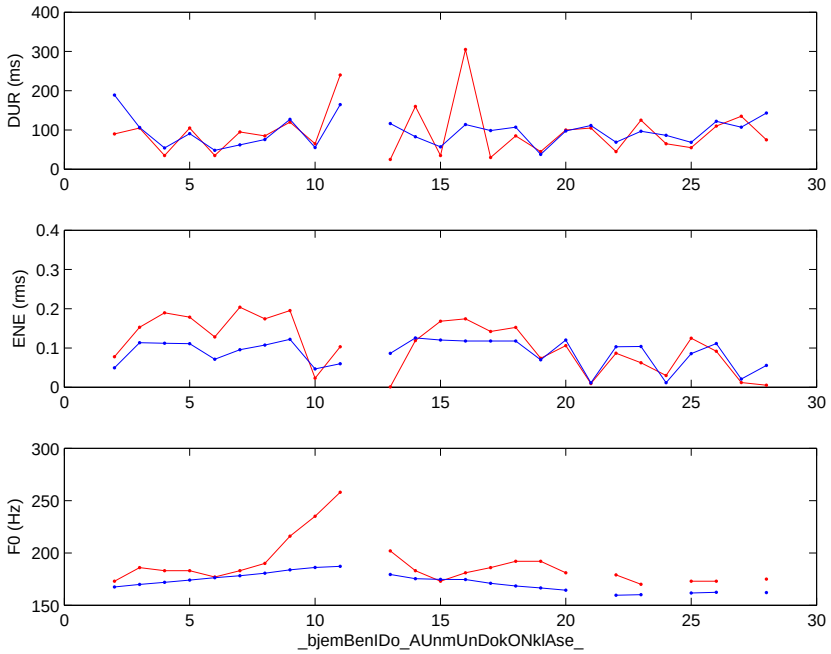


Figura D.75: Valores de duración (ms), energía (rms) y F_0 (Hz) calculados por el sistema (color azul) comparados con los de la misma frase del corpus (color rojo) para la frase núm. 15 del estilo triste.

D.6. Instrucciones de la prueba subjetiva

La prueba de percepción se presenta al participante en un entorno web. Una vez ha cumplimentado la página de acceso con su dirección de correo electrónico, aparece una página inicial que contiene un texto con las instrucciones y seis enunciados de ejemplo para que el oyente se familiarice con el habla del estilo que se pretende evaluar. Los textos con las instrucciones varían ligeramente del estilo neutro al resto, ya que se matiza que el estilo neutro no trata de transmitir ningún estado de ánimo en particular. Los textos presentados al oyente se muestran a continuación:

Estilo neutro:

A continuación escucharás una serie de frases creadas automáticamente por un ordenador en las que no se intenta reproducir ***ningún estado de ánimo en particular****. No se trata de evaluar si entiendes cada palabra, sino de que prestes atención a la frase entera, sin tener en cuenta algunas alteraciones en la pronunciación concreta que puedas encontrar en algunos casos. Puedes fijarte, por ejemplo, en* ***el tono en que está pronunciada****, en la* ***rapidez con la que se habla*** *o en la* ***fuerza*** *con la que se pronuncia, teniendo siempre en cuenta que se trata de una locución neutra que no intenta transmitir ningún estado de ánimo concreto.*

En esta página puedes escuchar unos ejemplos para familiarizarte con la voz sintética. A partir de la siguiente página empezará tu evaluación. Puedes abandonarla en cualquier momento y reanudarla accediendo con la misma dirección de e-mail.

Resto de estilos (ejemplo extraído del estilo alegre):

A continuación escucharás una serie de frases creadas automáticamente por un ordenador en las que se intenta reproducir un estado de ánimo ***ALEGRE****. Te pedimos que valores específicamente los aspectos relacionados con la manera en que cada frase transmite la emoción deseada, sin tener en cuenta algunas alteraciones en la pronunciación concreta de cada palabra que puedas encontrar en algunos casos. No se trata de evaluar si entiendes cada palabra, sino de que prestes atención a la frase entera. Puedes fijarte, por ejemplo, en* ***el tono en que está pronunciada*** *, en el* ***énfasis*** *en algunas partes o en toda la frase, en* ***la rapidez con la que se habla*** *o en* ***la fuerza con la que se pronuncia****, teniendo siempre en cuenta qué estado de ánimo se pretende transmitir.*

En esta página puedes escuchar unos ejemplos para familiarizarte con la voz sintética. A partir de la siguiente página empezará tu evaluación. Puedes abandonarla en cualquier momento y reanudarla accediendo con la misma dirección de e-mail.

El enunciado de la pregunta que se le presenta al participante en la evaluación de cada estímulo es de uno de los dos tipos siguientes en función de si se evalúa el estilo neutro o cualquier otro (p.ej. el estilo alegre):

Esta frase no pretende transmitir ningún estado de ánimo en particular. Consideras que su pronunciación global (tono, velocidad, acentuación) es:

Esta frase pretende transmitir alegría. Consideras que su pronunciación global (tono, velocidad, acentuación) es:

Excelente (5)
Buena (4)
Regular (3)
Mediocre (2)
Mala (1)

Finalmente, al participante se le pide que indique el sexo, la edad y se le da la posibilidad de añadir un comentario.

Apéndice E

Análisis del texto

E.1. SINLIB. Herramienta para el análisis del texto

En este apartado se presenta el módulo, basado en la generación e implementación de un lenguaje para la interpretación de reglas, que permite la conversión de un texto en su correspondiente transcripción fonética y, además, la asignación de propiedades relacionadas con los atributos prosódicos. Una descripción más detallada de la implementación y de la utilización de esta herramienta se puede encontrar en Sánchez (1997).

El sistema se ha implementado utilizando diagramas sintácticos que permiten la descripción del lenguaje que posibilita la programación de las reglas que se aplicarán al texto.

Se definió un lenguaje en función de los objetivos siguientes:

- Ofrecer un modo de realizar la conversión de grafema a fonema o alófono y la asignación de atributos prosódicos basado en reglas externas al código del programa.
- Conseguir unas reglas sencillas, claras y fácilmente modificables por el usuario.
- Permitir modificaciones en la sintaxis de las reglas.

E.1.1. Características del lenguaje

Las características principales del lenguaje desarrollado son:

- Las reglas se estructuran en módulos de reglas denominados MODR, de forma que puedan agruparse según su cometido. Por ejemplo: reglas de acentuación, reglas de transcripción fonética, etc.
- Los ficheros de reglas son ficheros de texto, de modo que el usuario pueda editarlos con facilidad.
- Las reglas se procesan secuencialmente del principio al final del MODR.
- Un fichero de comportamiento incluye los nombres y la ubicación de los diferentes ficheros de reglas, así como su orden de aplicación.
- Las reglas se compilarán previamente para garantizar que la sintaxis sea correcta.
- Las reglas actúan sobre una estructura de datos, que consiste, básicamente, en una lista de elementos que representan los grafemas/fonemas y alófonos que la frase que contiene.
- Las reglas tienen una estructura condición $\Rightarrow$ acción, la acción solo se llevará a cabo si la condición se cumple.

Podemos distinguir dos fases en el funcionamiento del sistema: la compilación de reglas (solo una vez o en el caso de modificar o añadir reglas) y la ejecución de reglas (cada vez que hay que realizar la conversión de texto a fonemas o alófonos).

El lenguaje intenta minimizar el número de funciones que emplea con la finalidad de mantener una baja complejidad. A continuación se describen las funciones de las que consta el lenguaje:

Ina (posItem, propiedad): Hace referencia a una propiedad de un ítem en concreto de la lista para que sea consultada o modificada. La regla puede aplicarse tanto en la parte de condición de la regla como en la parte de acción. Parámetros que recibe:

posItem: Posición del ítem a que hace referencia con respecto al índice actual; por ejemplo, posItem = 0 indica el ítem actual, posItem = 1 indica el ítem anterior y posItem = -1 el ítem posterior, siempre con respecto al ítem actual.

propiedad: Especifica la propiedad que se quiere consultar o actualizar (véase la lista de propiedades en la tabla E.2).

Ejemplo:

// Si el ítem actual es un grafema 'a' entonces la propiedad VOCAL es cierta.

Ina(0 , GRAFEMA) == 'a' ⇒ Ina(0 , VOCAL) = TRUE ;

EliminaItem (posItem): Elimina el ítem indicado. La regla sólo se aplica en la parte acción de la regla. Cuando se elimina un ítem, el ítem actual pasa a ser el ítem posterior al eliminado.

Parámetros que recibe: *posItem*

Ejemplo:

// Si el grafema actual es una 'h', entonces se elimina.

Ina(0, GRAFEMA) == 'h' ⇒ EliminaItem(0);

InsertaItem (posItem): Inserta un ítem en la lista. La regla solo se aplica en la parte acción de la regla. Cuando se inserta un ítem, el ítem actual pasa a ser el ítem recién insertado.

Parámetros que recibe: *posItem*

Ejemplo:

// Si el grafema es una 'x', se convierte en los fonemas 'k' y 's'.

Ina(0 , GRAFEMA) == 'x' ⇒ Ina(0,FONEMA)='k' , InsertaItem(0) , Ina(0,FONEMA)='s';

El *token* es la unidad mínima de representación en un lenguaje. En el caso del conversor implementado un *token* puede ser:

- Un elemento como un paréntesis, una coma, etc.
- Un número (entero o en coma flotante).
- Una palabra reservada del lenguaje.
- Un fonema o un alófono.
- Una palabra en sentido general.

En las tablas E.1 y E.2 se muestran las listas de *tokens* que admite el lenguaje y de las propiedades que se han definido, respectivamente.

Tabla E.1: Lista de *tokens*.

Token	Descripción
(	Paréntesis abierto. Se utiliza para el anidamiento de condiciones
)	Paréntesis cerrado. Se utiliza para el anidamiento de condiciones
==	Operador de comparación
⇒	Operador separador entre la parte condición y acción de la regla
!=	Operador diferente
=	Operador asignación
+=	Operador suma e igualación
−=	Operador resta e igualación
and	Operador lógico *and*
or	Operador lógico *or*
TRUE	Operador TRUE
FALSE	Operador FALSE
//	Comentario
,	Separador de funciones en la parte acción de la condición
’	Delimitador de fonema o grafema
;	Indicador de fin de regla
”	Delimitador de palabra

E.1.2. Módulos del sistema

Los módulos que forman el sistema implementado se describen a continuación:

Preprocesador. Se encarga de recibir el texto y crear la lista de elementos inicializada con el valor del grafema correspondiente; además, marca algunos *flags* de dentro de la estructura como los de inicio y fin de palabra.

Generador de salida. Se encarga de generar un fichero con toda la información que contiene la estructura de datos en un instante dado. Dicha estructura se puede cargar con posterioridad para recuperar la totalidad de la información contenida en la estructura.

Tabla E.2: Lista de propiedades

Nombre de la propiedad	Valores que toma	Descripción
PREPAUSAL	TRUE o FALSE	Pertenece a una sílaba prepausal
INI_SILABA	TRUE o FALSE	Si el ítem es inicio de sílaba
FIN_SILABA	TRUE o FALSE	Si el ítem es final de sílaba
VOCAL	TRUE o FALSE	Si es vocal (true) o consonante (false)
GRUP_CONS	TRUE o FALSE	Pertenece a un grupo consonántico
INI_PALABRA	TRUE o FALSE	Si es comienzo de palabra
FIN_PALABRA	TRUE o FALSE	Si es final de palabra
ACENTO_GRAFICO	TRUE o FALSE	Si el ítem tiene acento gráfico
TRABADA	TRUE o FALSE	Si la vocal está en posición trabada
ACENTO	TRUE o FALSE	Si el ítem es una vocal acentuada
VOC_FINAL	TRUE o FALSE	Si es vocal final de palabra
EXCEPCION	TRUE o FALSE	La palabra actual es una excepción
GRAFEMA	Identificador	Valor del grafema del ítem actual
FONEMA	Identificador	Valor del fonema o alófono del ítem actual
PALABRA_ACT	String	Palabra a la cual pertenece el ítem actual
PALABRA_ACENT	TRUE o FALSE	Si la palabra actual está acentuada
ULTIMA_SILABA	TRUE o FALSE	Si es la última sílaba de la palabra
PENULTIMA_SILABA	TRUE o FALSE	Si es la penúltima sílaba de la palabra

Ejecutador de MODR. Se encarga de ir ejecutando módulos de reglas según lo especificado mediante el fichero en el cual se describe el comportamiento del sistema.

Intérprete del lenguaje de comportamiento. Realiza un *parsing* de la información contenida en el fichero de descripción de comportamiento y guarda una representación interna de dicha información.

Estructura de datos. La estructura de datos contiene una representación de la información que trata el sistema. La estructura está orientada hacia un párrafo de texto, el cual está a su vez compuesto de elementos que pueden ser grafemas (antes de que se haya procesado la información) o fonemas/alófonos (después de que se haya procesado).

Gestión. El módulo de gestión se encarga de coordinar las acciones del *Parser*, del *Scanner* y del Ejecutador.

Scanner. Lee el texto de las reglas y pasa una serie de *tokens* al *Parser*.

Parser. Verifica que la sintaxis de las reglas sea correcta. Genera un código intermedio (*p-code*) el cual se pasa posteriormente al Ejecutador para que sea procesado.

Universitat Ramon Llull

Aquesta Tesi Doctoral ha estat defensada el dia ____ d _____ de 200

al Centre Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle

de la Universitat Ramon Llull

davant el Tribunal format pels Doctors sotasignants, havent obtingut la qualificació:

President/a

_____

Vocal

_____

Vocal

_____

Vocal

_____

Secretari/ària

_____

Doctorand/a
Ignacio Iriondo Sanz

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es