eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TESIS DOCTORAL

Análisis y Evaluación de Parámetros para Identificación Automática de Emociones en el Habla

Iker Luengo

Aholab Signal Processing Laboratory
Departamento de Electrónica y Telecomunicaciones
Universidad del País Vaso / Euskal Herriko Unibertsitatea

Bilbao, 2010

A mi familia.
A los que están y a los que ya se han ido.

Resumen

LOS sistemas de identificación automática de emociones en la voz tratan de determinar el estado emocional de un locutor a partir de las características de su voz. El número y tipo de características utilizadas son, por tanto, uno de los elementos clave a la hora de diseñar este tipo de sistemas. Es necesario que los parámetros utilizados caractericen adecuadamente los diferentes estilos de habla, permitiendo una correcta identificación de cada emoción y minimizando la probabilidad de confusión entre estas emociones.

Sin embargo, no existe un consenso sobre qué parámetros son los mejores para la identificación de emociones. Un repaso de la literatura permite comprobar cómo los grupos de investigación utilizan diferentes conjuntos de parámetros, muy diferentes entre sí. Se utilizan indistintamente parámetros derivados de la forma espectral de la señal, las características prosódicas, la señal glotal o incluso valores derivados del contenido lingüístico, en un intento de hallar aquellos parámetros que realmente sean útiles.

Este fenómeno se agrava por la falta de un estudio sistemático que analice la efectividad de cada parametrización con el objetivo de determinar la más favorable. En la literatura existen varios trabajos que analizan la capacidad proporcionada por diferentes parámetros para la discriminación de emociones. Sin embargo, no presentan una visión completa del comportamiento de estas parametrizaciones. Muchos de estos trabajos analizan cada parámetro de forma individual, lo que puede dar lugar a interpretaciones no válidas cuando estos parámetros se utilizan en combinación con otros. Otros estudios proporcionan tasas de precisión obtenidas mediante evaluación experimental para una cierta parametrización, con lo que se puede deducir el comportamiento global de todo el conjunto de parámetros. Sin embargo, no se suelen proporcionar los resultados de las parametrizaciones por separado, dando sólo los resultados de la combinación total, con lo que no es posible saber si esta combinación realmente es mejor que las partes por separado. Además, los diferentes estudios presentes en la literatura no son comparables entre sí, debido a las diferencias en la arquitectura de los experimentos, el número y tipo de emociones, la calidad de las grabaciones u otros factores. Por lo tanto, no es posible construir una visión completa de las propiedades de cada parametrización

simplemente comparando los resultados de los diferentes trabajos publicados.

Mediante el trabajo desarrollado en esta tesis se pretende cubrir este vacío del campo del habla emocionada. Se presenta un análisis sistemático de las parametrizaciones acústicas más comúnmente utilizadas en la identificación automática de emociones, determinando así su capacidad para distinguir los diferentes estilos de habla y su efectividad en los sistemas de identificación. Se ha tenido especial cuidado en lograr que los resultados obtenidos para cada una de estas parametrizaciones sean comparables entre sí, utilizando para ello bases de datos y arquitecturas comunes durante todo el proceso.

El análisis presentado en este documento se ha llevado a cabo mediante diferentes métodos, estudiando cada parámetro por separado (a través de técnicas de ranking de parámetros), así como considerando todo el conjunto de parámetros (mediante medidas de dispersión multidimensional). También se describen los resultados obtenidos en pruebas experimentales de identificación automática, lo que permite validar las conclusiones obtenidas durante el análisis.

En una primera fase, se ha realizado el análisis sobre emociones actuadas, utilizando la base de datos de habla emocional *Berlin*. Las conclusiones resultantes han sido posteriormente validadas en emociones naturales y habla espontánea, mediante la base de datos *AIBO*. En ambos casos se ha llegado a conclusiones similares, mostrando que aquellas parametrizaciones que destacan por su capacidad de discriminar emociones actuadas también obtienen los mejores resultados en situaciones más reales.

Los resultados desvelan que los parámetros prosódicos o de calidad de voz más habitualmente utilizados no son los más adecuados para la identificación automática de emociones, ya que las características espectrales presentan mayor capacidad de discriminación. Este efecto es más acusado cuando se consideran emociones naturales en habla espontánea.

Un análisis detallado de los parámetros sugiere que el escaso rendimiento de los parámetros prosódicos y de calidad de voz probablemente se debe a la dificultad en el cálculo de estos parámetros a partir de la señal de voz. Esta dificultad es todavía más evidente en voz espontánea. Aunque las características prosódicas y de calidad de voz son generalmente consideradas como un vehículo importante de la información emocional, la extracción de esta información mediante algoritmos automáticos no es sencilla, lo que provoca que los parámetros estimados sean poco robustos y aumenten la confusión en el sistema de clasificación. Por el contrario, las características espectrales muestran mayor estabilidad, a la vez que transportan una cantidad de información emocional considerable, haciendo que finalmente presenten mejores cualidades para la identificación.

Laburpena

AHOTSAREN bidez emozioak automatikoki ezagutzeko sistemek ahotsaren ezaugarrien bidez hizlari baten emozio-egoera zehaztea dute helburutzat. Beraz, halako sistemen diseinuaren gakoetako bat da ebaztea zein den erabilitako ezaugarrien mota eta kopurua. Ezaugarri horiek ondo bereizi behar dituzte hizketaren estilo ezberdinak, emozio bakoitza ondo ezagutzeko aukera emateko eta emozio horien arteko nahasketa murrizteko.

Hala ere, ez dago adostasunik emozioak ezagutzeko zein diren parametrorik onenak. Literatura gainbegiratuz gero, erraz ikusten da ikerketa-talde bakoitzak oso ezaugarri desberdinak erabiltzen dituela. Nahasturik erabiltzen dira seinalearen espektrotik, ezaugarri prosodikoetatik, ahots-korden seinaletik zein eduki linguistikotik eratorritako parametroak, benetan erabilgarriak diren ezaugarriak aurkitzeko ahaleginean.

Egoera hori are larriagotzen du parametrizaziorik onena zein den zehazten duen azterketa sistematiko bat ez izateak. Parametro ezberdinek emozioak bereizteko duten gaitasuna aztertzen duten hainbat lan ageri dira literaturan. Hala ere, ez dute parametrizazioen berezitasunen ikuspegi osoa ematen. Lan horietako askok banaka aztertzen dituzte parametroak, eta litekeena da ateratako ondorioak baliogabeak izatea parametrook bereiz edo beste batzuekin batera erabiltzen direnean. Beste lan batzuek ebaluazio esperimentalen bidez lortutako zehaztasun tasak ematen dituzte, parametrizazio jakin baterako. Horrela, ezaugarri multzo baten jokaera globala jakin daiteke. Baina, eskuarki, parametroen konbinazio osoaren emaitza ematen da soilik, eta ez azpi-parametrizazio ezberdinek banaka ematen dutena. Beraz, ezin jakin daiteke konbinazio batek bere azpi-parametrizazioek baino zehaztasun hobea ematen duen ala ez. Halaber, literaturan aurkeztutako ikerketak ezin dira bata bestearekin alderatu, esperimentuen arkitektura, emozioen mota eta kopurua, grabaketen kalitatea eta beste hainbat ezberdintasun direla eta. Beraz, ezin daiteke eraiki, argitaratutako lan ezberdinen emaitzak konparatuz, parametrizazio bakoitzaren berezitasunen ikuspegi osoa.

Tesi honetan garatutako lanarekin, hizketa emozionalaren arloan dagoen hutsune hori bete nahi izan da: emozioen ezagutza automatikoan gehien erabilitako parametrizazio akustikoen azterketa sistematikoa egin da, eta, hala, parametriza-

zio bakoitzak hizketa-estilo ezberdinak bereizteko duen gaitasuna ondorioztatu da. Kontu handia izan da parametrizazio bakoitzarekin lortutako emaitzak bata bestearekin konparagarriak izan daitezen, eta, horretarako, datu-base eta arkitektura berdinak erabili dira prozesu osoan.

Dokumentu honetan aurkeztutako analisia hainbat metodoren bidez gauzatu da: alde batetik, parametro bakoitza bakarka ikertuz (parametroak sailkatzeko teknikak erabiliz); bestetik, parametroen multzo osoa kontuan hartuz (banaketen sakabanaketaren neurketak eginez). Emozioak automatikoki ezagutzeko esperimentuen emaitzak ere eman dira, horien bidez analisian lortutako ondorioak egiaztatzeko.

Lehenengo atalean, antzeztutako emozioekin egin da analisia, *Berlin* datubasea erabilita. Gero, lortutako ondorioak emozio naturaletarako eta bat-bateko hizketarako egiaztatu dira, *AIBO* datu-basea erabiliz. Kasu bietan antzeko ondorioetara heldu da: antzeztutako emozioak ondo bereizteko gai diren parametroek emaitza hoberenak ematen dituzte egoera errealetan ere.

Lortutako emaitzen bidez ikus daiteke ezaugarri prosodikoetatik edo ahots-kalitatetik eratorritako parametroak ez direla onenak emozioak automatikoki ezagutzeko, nahiz literaturan erabilienak izan. Ahotsaren ezaugarri espektralek emozioak bereizteko gaitasun handiagoa dute; batez ere, emozio naturalak eta bat-bateko hizketa baliatzen direnean.

Parametroak xehe aztertuta, ondorioztatu da ezaugarri prosodikoetatik eta ahots-kalitatetik lortutako ezaugarrien etekin eskasa ezaugarri horien kalkuluaren zailtasunean datzala. Zailtasun hori are handiagoa da bat-bateko hizkeran. Nahiz eta aski hedatua eta onartua den prosodiaren eta ahots-kalitatearen ezaugarriak emozio-informazioaren garraio garrantzitsua direla, ez da erraza algoritmo automatikoen bidez informazio hori ateratzea; horrenbestez, lortutako parametroak ez dira oso sendoak, eta nahasmena handiagotzen dute ezagutza-sisteman. Ezaugarri espektralak, bestalde, egonkorragoak dira, eta aldi berean emozioari buruzko informazio nahikoa daramate; hortaz, aproposagoak dira ezagutzarako.

Abstract

AUTOMATIC systems for emotion identification in the speech seek to detect the emotional state of a speaker analysing the characteristics of his or her voice. Therefore, the type and number of characteristics that are going to be used is one of the key factors during the design of this kind of systems. The features should characterise the considered speaking styles adequately, so that they enable a correct identification and minimise the confusion probability among emotions.

Nevertheless, there is no consensus about which features are best for the emotion identification. A review of the literature shows that research groups use diverse feature sets, very different one from another. Parameters derived from the spectrum, prosodic characteristics, glottal pulse or linguistic content are used indiscriminately, in an attempt to capture those features that are really useful.

This phenomenon is aggravated due to the lack of a systematic study of the effectiveness of each parametrisation. There are several works in the literature that analyse the capability of different features to discriminate emotions. However, they do not provide a complete view of the behaviour of these parametrisations. Many of these works analyse each feature individually, which may lead to conclusions that are not accurate when these features are used in combination with others. Some other works provide experimental accuracy rates for a given parametrisation, so that the behaviour of the whole feature set can be deduced. However, only the accuracy of the complete feature set is usually given, and not the results of the different sub-parametrisations. Therefore, it is not possible to know wether this combination is really better than the individual sub-parametrisations or not. Furthermore, the studies present in the literature are not usually comparable among them, due to the differences in the architecture of the experiments, the number and type of emotions, the quality of the recordings, or many other factors. Therefore, it is not possible to build a complete view of the properties of each parametrisation simply comparing the results from different works that have been published.

The work developed in this thesis tries to cover this gap in the field of emotional speech. It presents a systematic analysis of the acoustic parametrisations that are commonly used in the automatic identification of emotions, evaluating

their capability to distinguish the different speaking styles and their effectiveness in the automatic identification systems. A special care has been taken in order to ensure that the results are comparable among all the considered parametrisations. For that purpose, common databases and architectures have been used all along the process.

The analysis presented in this document has been carried out with different methods, evaluating each feature individually (by means of feature ranking techniques), but also considering the complete feature set (by means of multidimensional dispersion measures). Results of automatic emotion identification experiments are also described, so that the conclusions derived from the analysis step can be confirmed.

A first evaluation has been done with acted emotions, using the *Berlin* emotional database. The resulting conclusions have been validated afterwards for natural emotions and spontaneous speech, with the *AIBO* database. In both cases the results have been similar, showing that the parametrisations that provide good discrimination for acted emotions are also the ones that get best results in real conditions.

Results reveal that prosodic or voice quality features, which are the most used ones for this task, are not the best choice for automatic emotion identification in the speech, and that spectral features have better discrimination capability. This effect is more noticeable with natural emotions and spontaneous speech.

A detailed analysis of the features suggests that the poor performance of prosodic and voice quality parameters is probably due to the difficulty of estimating these features directly from the voice signal. This difficulty is more evident in the case of spontaneous speech. Although it is considered that prosodic and voice quality characteristics carry most of the emotional information, it is very difficult to extract this information with automatic algorithms. This causes the estimated features to lack robustness and to increase the confusion in the identification system. On the contrary, spectral parameters show greater stability, and at the same time, they carry a respectable amount of emotional information, making them more suitable for identification purposes.

Agradecimientos

FINALIZADA la tesis, me enfrento al último problema, el de agradecer a todos aquellos que la han hecho posible, sin dejar a nadie sin mencionar.

En primer lugar, quiero expresar mi más sincero agradecimiento a mi directora de tesis, Eva Navas, por su inestimable ayuda. Por haber compartido conmigo las alegrías y las tristezas de este trabajo, y por haberme ayudado a ver un rayo de esperanza cuando parecía que no había salida. También debo agradecerle las múltiples revisiones que ha realizado de este documento, y las correcciones propuestas. Sólo espero que al menos hayan hecho menos aburridos todos esos viajes de autobús.

Asímismo, no puedo dejar de dar las gracias a Inma Hernáez, responsable del *AhoLab Signal Processing Laboratory*, por la confianza y el apoyo incondicional que me ha mostrado desde el primer día que entré en este grupo. A ella le debo la experiencia adquirida en todos estos años de investigación, sin los cuales esta tesis no hubiera podido ver la luz.

También estoy en deuda con mis compañeros de *AhoLab*. Con Iñaki Sainz y Jon Bonilla, por resolver con prontitud los problemas con el servidor. Con Jon Sánchez, por llevarme a tomar café y levantarme el ánimo siempre que me veía demasiado estresado. Con Ibon Saratxaga, por hacerme ver los problemas desde otro punto de vista y proponer soluciones alternativas. A Igor Odriozola debo agradecerle el estar siempre dispuesto a revisar mis documentos en euskera. Eskerrik asko. También he de mencionar a Daniel, Eneritz, Amaia y Nora. Habéis sido todos unos compañeros estupendos.

Por último, quiero expresar mi más profundo agradecimiento a mi familia, en especial a mi padre y a mi madre. Aunque los haya dejado para el final, no significa que sean menos importantes. Puedo decir, sin temor a equivocarme, que han sufrido esta tesis tanto como yo. Durante su ejecución, así como en el resto de las etapas de mi vida, han estado ahí, dándome ánimos y empujándome hacia delante cuando las cosas se torcían. Por todo ello, gracias. Siempre estaré en deuda con vosotros.

Índice general

Índice de figuras

Índice de tablas

Acrónimos y abreviaturas

ANN Red neuronal artificial
(*Artificial Neural Network*)

apq5 Cociente de perturbación de amplitud de cinco puntos
(*five-point Amplitude Perturbation Quotient*)

CDP Cepstrum y programación dinámica
(*Cepstrum and Dynamic Programming*)

CMS Normalización de la media cepstral
(*Cepstral Mean Subtraction*)

CTH Conversión de texto a habla

CQ Cociente de cierre
(*Closing Quotient*)

DCT Transformada coseno discreta
(*Discrete Cosine Transform*)

DFT Transformada discreta de Fourier
(*Discrete Fourier Transform*)

fdp Función de densidad de probabilidad

GMM Modelo de mezcla de gaussianas
(*Gaussian Mixture Model*)

HMM Modelo oculto de Markov
(*Hidden Markov Model*)

IAIF Filtrado inverso adaptativo iterativo
(*Iterative Adaptative Inverse Filtering*)

IDFT Transformada discreta de Fourier inversa
(*Inverse Discrete Fourier Transform*)

kNN *k* vecinos más próximos
(*k-Nearest Neighbors*)

LDA Análisis lineal discriminante
(*Linear Discriminant Analysis*)

LFPC Coeficientes de potencia en escala logarítmica
(*Log-Frequency Power Coefficients*)

LPC Coeficientes de predicción lineal
(*Linear Prediction Coefficients*)

LPCC Coeficientes cepstrales de predicción lineal
(*Linear Prediction Cepstral Coefficients*)

LTSD Divergencia espectral a largo plazo
(*Long-Term Spectral Divergence*)

LTSE Envolvente espectral a largo plazo
(*Long-Term Spectral Envelope*)

MFCC Coeficientes cepstrales en escala mel
(*Mel Frquency Cepstral Coefficients*)

mRMR Mínima-redundancia-máxima-relevancia
(*Minimal-Redundancy-Maximal-Relevance*)

NAQ Cociente de amplitud normalizada
(*Normalized Amplitude Quotient*)

ppq5 Cociente de perturbación de período de cinco puntos
(*Five-point Period Perturbation Quotient*)

RAE Reconocimiento automático de emociones

RAH Reconocimiento automático del habla

RMSE Raíz del error cuadrático medio
(*Root Mean Square Error*)

SB Equilibrio espectral
(*Spectral Balance*)

SNR Relación señal a ruido
(*Signal to Noise Ratio*)

SVM Máquina de vectores soporte
(*Support Vector Machine*)

UAR Precisión media no ponderada
(*Unweighted Average Recall*)

VAD Detección de actividad vocal
(*Voice Activity Detection*)

VQ Calidad de voz
(*Voice Quality*)

VUV Sordo-sonoro
(*Voiced-UnVoiced*)

WAR Precisión media ponderada
(*Weighted Average Recall*)

WOZ Mago de Oz
(*Wizard of OZ*)

Capítulo 1

Introducción

Índice

LAS emociones forman parte de nuestra vida cotidiana. Todos los días sentimos emociones que asumimos como parte de nuestra rutina de forma natural: nerviosismo, alegría, tedio, melancolía, excitación, vergüenza, exasperación, abatimiento... La lengua contiene una gran cantidad de términos para expresar estados emocionales. Según Cowie (2000) esta riqueza de vocabulario para referirse a estados emocionales es debida a la importancia que las mismas tienen sobre nuestra vida, de forma que buscamos expresar con palabras el gran abanico de emociones que podemos distinguir, algunas de ellas con diferencias muy sutiles entre sí.

La importancia que las emociones tienen sobre el ser humano radica en su función como herramienta de comunicación entre las personas. Los sentimientos no son un elemento únicamente propio y personal, sino que traspasa las fronteras del individuo llegando a ser una herramienta para establecer relaciones sociales. Los seres humanos somos capaces de percibir el estado emocional de nuestros semejantes y esta capacidad proporciona una importante fuente de información y comunicación. De esta forma somos capaces de modificar, consciente o inconscientemente, nuestro propio comportamiento en función de las emociones percibidas en los demás. Por ejemplo podemos disculparnos si vemos que hemos molestado a alguien, modificar nuestro discurso si el interlocutor se aburre o esforzarnos más si detectamos que nuestros superiores se muestran decepcionados con nuestro trabajo.

No sólo respondemos al estado emocional del prójimo, sino que muchas veces confiamos en que nuestras propias emociones se expongan de forma clara ante los demás sin necesidad de tener que expresarlas con palabras. Muchas veces basta con mirar con dureza para enviar un completo mensaje de desaprobación, o sonreír para comunicar todo lo contrario. Esta comunicación no verbal de las emociones está tan arraigada en nuestra forma natural de relacionarnos con los demás que la utilizamos incluso para relacionarnos con animales y objetos. Es habitual que una persona hable con su mascota tal y como lo haría con un niño, premiándolo o reprendiéndolo con el propio tono de voz. Pero no es menos habitual ver a alguien gritando a la pantalla de su ordenador cuando no consigue hacerlo funcionar correctamente. Reeves y Clifford (2003) presentan una serie de experimentos en los

que se comprueba que los humanos tendemos a tratar con las máquinas como si fueran personas, sin importanos que estas máquinas sean incapaces de comprender nuestros sentimientos.

En una época en la que se han hecho grandes avances para conseguir una comunicación oral con las máquinas, desarrollando sistemas de reconocimiento automático del habla y de síntesis, esta comunicación sigue considerándose *fría*, no del todo natural. Gran parte de la culpa de esta percepción la tiene precisamente el hecho de que las máquinas no son empáticas, no pueden detectar nuestras emociones y modificar su comportamiento de acuerdo a ellas. Por tanto, la comunicación no es totalmente natural, tal y como se daría con otro ser humano.

Una vez que se ha implementado una interfaz de comunicación verbal básica con las máquinas (es decir, que puedan entender un mensaje oral y responder también de forma oral), el siguiente paso para conseguir una mayor naturalidad consiste indudablemente en conseguir que sean capaces de entender y atender a las expresiones no verbales que comunican nuestro estado de ánimo, y por qué no, expresar también emociones en su respuesta. Aunque las emociones expresadas por la máquina no sean auténticas, no las *sienta*, sí proporcionan una naturalidad añadida al mensaje, que puede hacer que un interlocutor humano considere la comunicación más agradable.

1.1. Aplicaciones de la identificación de emociones

Un sistema de identificación automática de emociones proporciona la herramienta necesaria para que, con las reglas adecuadas, una máquina pueda tomar decisiones y modificar su comportamiento en función de la emoción detectada. Esto haría que la comunicación hombre-máquina fuera más fluida y percibida con mayor naturalidad por parte de las personas, que verían que la máquina responde a sus mensajes no verbales.

Sin embargo, sin llegar a buscar una comunicación emocional completamente natural con las máquinas (lo que de momento parece estar más allá de las posibilidades de la tecnología), existen aplicaciones inmediatas que pueden beneficiarse de la detección automática de emociones. Una de las primeras que han sido desarrolladas consiste en la detección de personas enfadadas o frustradas en sistemas automáticos de atención al cliente. Esta detección permite, por ejemplo, transferir la llamada a un operador humano, que posiblemente esté más capacitado para entender y resolver el problema del cliente, así como para intentar calmarlo. De esta forma se evita tener clientes insatisfechos a causa de una máquina que no puede resolver sus problemas (Kim *et al.*, 2007; Petrushin, 2000).

Este sistema también permite implementar un control automático de la calidad del servicio ofrecido por el sistema automático de atención. Si el número de clien-

tes que no pueden comunicarse el sistema es excesivo, puede que este sistema no sea rentable. Claro que esto también puede aplicarse a operadores humanos: si las llamadas atendidas por un operador humano en particular acaban frecuentemente con clientes enfadados, es posible que la atención ofrecida por ese operador no sea la adecuada.

Si en lugar de centrarnos en detectar el enfado tratamos de detectar el estrés y el nerviosismo en la voz de una persona, podemos utilizar el sistema en un entorno diferente. Midiendo el nivel de estrés y angustia se puede realizar una estimación de la urgencia de una llamada a un servicio de emergencias y priorizar la atención. Un sistema similar ya ha sido propuesto por Petrushin (2000) para un sistema de mensajería de voz (*voice mail*), donde los mensajes se ordenan por prioridad según la urgencia expresada por la voz del locutor.

Desde un punto de vista más lúdico, se pueden desarrollar juguetes y mascotas mecánicas con las que se pueda interaccionar emocionalmente de forma natural, como con un perro que sabe cuándo su amo está triste o le está reprendiendo. Ya existen en el mercado mascotas mecánicas programadas para *sentir* diferentes emociones según las circunstancias. Algunos ejemplos son Furby (Tiger Electronics), Aibo (Sony) o Qrio (Sony). El siguiente paso es conseguir que esta comunicación emocional sea bidireccional, es decir, que también identifique las emociones de sus amos. En este sentido se están desarrollando varios trabajos de investigación, como por ejemplo el de Dornaikaa y Raducanub (2007), que han implementado un sistema de identificación de emociones en rostros y lo han integrado en el robot Aibo. Aunque de momento la única respuesta del robot ante una emoción es el imitarla, supone una herramienta fundamental para que en un futuro el robot pueda reaccionar en función de la emoción detectada.

Si nos centramos en la tecnología del habla, una correcta identificación de la emoción en la voz puede permitir mejorar la precisión de los sistemas de reconocimiento automático del habla y de reconocimiento de locutor. Los sistemas de reconocimiento están generalmente entrenados con voz neutra, no emocionada. Sin embargo, es sabido que el estado emocional de un locutor afecta al sistema fonador, y por tanto, a las características de la voz. Debido a esto, los sistemas de reconocimiento presentan una mayor tasa de errores cuando son utilizados con voz emocionada (Bosch, 2003). Si fuéramos capaces de detectar esta emoción, se podría adaptar el sistema de reconocimiento para contrarrestar su efecto, bien adaptando los modelos utilizados, o bien realizando algún tipo de normalización como proponen Wu *et al.* (2006).

Otra aplicación interesante en el campo de las tecnologías del habla es la posibilidad de realizar una transcripción automática de las emociones de un locutor. Junto con un sistema de reconocimiento automático del habla esto permitiría no sólo saber qué se ha dicho, sino cómo se ha dicho. Por un lado, puede ser beneficioso para los sistemas de transcripción automática que se están desarrollando

para personas con deficiencias auditivas. Por otro, los sistemas de traducción automática voz-voz pueden utilizar esta información emocional para sintetizar la señal en el idioma de salida, imprimiendo a la voz una emoción similar a la detectada, de forma que el oyente percibe la totalidad del mensaje que el locutor pretendía transmitir, tanto de forma verbal como no verbal.

Por último, también existen aplicaciones enfocadas al aprendizaje y la educación emocional. Al igual que se están utilizando tecnologías de reconocimiento del habla para desarrollar sistemas para ayudar a que las personas con deficiencias auditivas y neuromusculares aprendan a vocalizar correctamente (Saz *et al.*, 2009), pueden desarrollarse programas que, a modo de juego o competición, permitan aprender a identificar las emociones del prójimo, tal y como propone hacer Petrushin (2000) con niños autistas.

1.2. Motivación y objetivos

Una de las claves a la hora de implementar un sistema de identificación automática de emociones en el habla es la selección de un conjunto de parámetros adecuado para esta tarea. Estos parámetros, utilizados para caracterizar los diferentes estilos de habla, tienen un gran efecto sobre los resultados del sistema de identificación. La fiabilidad de la clasificación, la capacidad para adaptarse a voces de locutores diferentes a los del entrenamiento, o para tratar con emociones en el habla natural dependen en gran medida del tipo de parametrización utilizada.

Sin embargo, no existe un consenso sobre qué parámetros son los mejores para la identificación de emociones. Un repaso de la literatura permite comprobar cómo los grupos de investigación utilizan diferentes conjuntos de parámetros, muy diferentes entre sí. Se utilizan indistintamente parámetros derivados de la forma espectral de la señal, las características prosódicas, la señal glotal o incluso valores derivados del contenido lingüístico, en un intento de hallar aquellos parámetros que realmente sean útiles. Este fenómeno se agrava por la falta de un estudio sistemático que analice la efectividad de cada parametrización con el objetivo de determinar la más favorable.

Este trabajo pretende cubrir, al menos en parte, este vacío. Por ello, el objetivo principal de esta tesis es realizar un análisis sistemático de las parametrizaciones más habitualmente utilizadas en los sistemas de identificación automática de emociones en el habla, y determinar su utilidad para esta identificación. Concretamente, el estudio se centra en los parámetros de naturaleza acústica: características espectrales, prosódicas y de calidad de voz.

Para alcanzar este objetivo y conseguir que las conclusiones obtenidas sean de utilidad en sistemas reales de identificación de emociones, es necesario, además, tener en cuenta los siguientes objetivos parciales:

- Realizar el análisis de tal forma que los resultados puedan ser comparados en igualdad de condiciones entre las diferentes parametrizaciones consideradas. Esto implica utilizar una misma base de datos y una misma metodología en todos los casos.

- Analizar no sólo las parametrizaciones individuales, sino sus combinaciones. Esto es necesario para poder tener una visión global de la eficacia de todas las posibilidades, puesto que la utilización conjunta de varias fuentes de información puede ser provechosa para reducir el error de identificación.

- Comparar y evaluar las dos técnicas de fusión de información más habituales a la hora de realizar la combinación de parametrizaciones: fusión temprana (o fusión de parámetros) y fusión tardía (o fusión de clasificadores).

- Determinar, a partir de los anteriores análisis, la capacidad de discriminación de emociones proporcionada por cada una de las parametrizaciones y sus combinaciones, y en última instancia, determinar cuál es la mejor.

- Realizar todos los procesos de parametrización de forma totalmente automática, sin supervisión humana. Generalmente, en una aplicación real de identificación automática de emociones, todo el proceso de extracción de parámetros debe realizarse de forma automática. Para que las conclusiones extraídas a lo largo de este trabajo sean aplicables a este tipo de entornos, es necesario que estas parametrizaciones se realicen bajo las mismas condiciones. Esto puede implicar tener que modificar algunos algoritmos de procesado de señal habitualmente utilizados durante la extracción de parámetros, o incluso diseñar algoritmos completamente nuevos.

1.3. Esquema de la tesis

El capítulo 2 realiza una revisión de la literatura publicada en el campo de la identificación automática de emociones en el habla. Presenta las alternativas más utilizadas para cada uno de los aspectos relevantes para el diseño de sistemas automáticos de identificación de emociones: bases de datos de desarrollo, parametrización de señales y algoritmos de clasificación. También recoge y compara la precisión obtenida por los trabajos de investigación más representativos en este campo, relacionando estos resultados con las técnicas utilizadas en cada caso.

El capítulo 3 describe los diferentes parámetros que van a ser evaluados y el procedimiento utilizado para su cálculo. Algunos de estos procedimientos han sido desarrollados especialmente para el estudio presentado en esta tesis. Todos los parámetros considerados (características espectrales, prosódicas y de calidad

de voz) tienen naturaleza acústica, es decir, pueden ser obtenidos directamente a partir de la propia señal de voz.

El capítulo 4 presenta el análisis de los parámetros en cuanto a su capacidad para discriminar emociones en la voz. Para el desarrollo de este análisis se utiliza una base de datos de habla emocional actuada que incluye grabaciones en siete estados emocionales diferentes. Los resultados son validados mediante pruebas empíricas de identificación automática de emociones.

En el capítulo 5 se comprueba si las conclusiones obtenidas en el análisis de los parámetros son generalizables. Para ello se realizan una serie de pruebas sobre una base de datos diferente. Esta nueva base de datos contiene grabaciones de cinco emociones naturales y un mayor número de locutores. Con estas pruebas se pretende comprobar si las conclusiones obtenidas acerca de la capacidad de discriminación de los parámetros pueden extrapolarse a otras emociones no consideras, así como al caso de emociones naturales y voz espontánea.

Por último, el capítulo 6 contiene las conclusiones generales derivadas de este estudio y alternativas para trabajos futuros.

Capítulo 2

Estado del arte en la identificación de emociones

Índice

LA identificación de emociones consiste en asignar una etiqueta emocional a una muestra de voz en función de la emoción percibida en la misma. En este sentido la identificación de emociones puede describirse como un problema de clasificación, donde cada muestra de voz es clasificada en una de las emociones consideradas.

La Figura 2.1 presenta la arquitectura típica de un sistema automático de clasificación, separando cada una de sus etapas. El sistema consta de dos fases, una de entrenamiento, en donde se crean los modelos de habla emocionada, y otra de clasificación, en donde se utilizan estos modelos para clasificar las muestras de entrada.

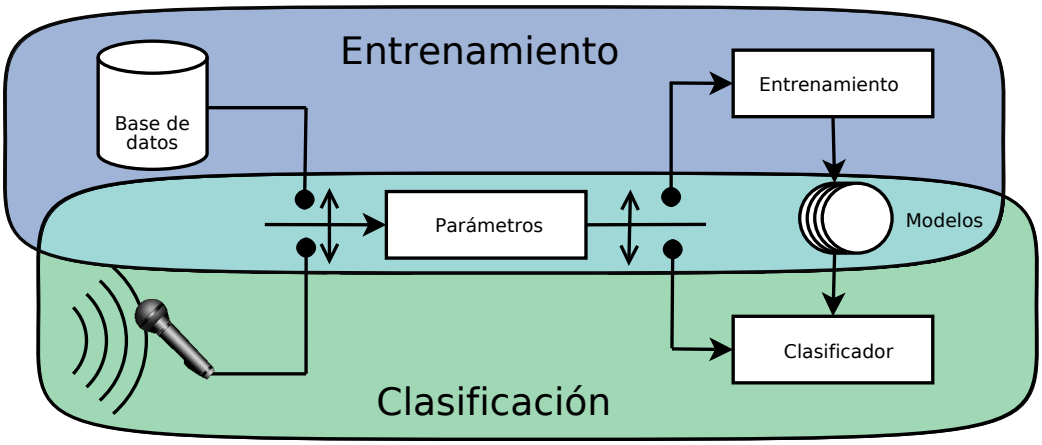


FIGURA 2.1: *Esquema general de un identificador de emociones.*

Durante la fase de entrenamiento el sistema aprende las características de cada estilo utilizando una base de datos de habla emocionada. Las señales de la base de datos se procesan para extraer parámetros que permitan distinguir las diferentes emociones entre sí. Con estos parámetros y conociendo la emoción a la que pertenece cada señal se entrenan los modelos del clasificador.

En la fase de clasificación el sistema se utiliza para identificar la emoción presente en muestras de voz desconocidas. Para ello se calculan los parámetros de la

nueva muestra, y se utilizan los modelos entrenados anteriormente para determinar su emoción.

En esta breve descripción del sistema de clasificación pueden identificarse los elementos clave para desarrollar el detector de emociones. La **base de datos de entrenamiento** define qué emociones será capaz de discriminar el sistema y bajo qué condiciones. Los **parámetros** deben ser cuidadosamente seleccionados puesto que determinan la capacidad del sistema para distinguir una emoción de otra. Por último, utilizar el tipo de **modelo** correcto es fundamental, ya que es el elemento del que depende la decisión final.

En las siguientes secciones de este capítulo se realiza una revisión de las alternativas que ofrece la literatura para cada uno de estos elementos clave. La sección 2.1 presenta las diferentes bases de datos de habla emocional disponibles, con cada una de sus características. Los diferentes parámetros utilizados para la identificación de emociones en el habla son tratados en la sección 2.2. En la sección 2.3 se analizan los posibles clasificadores y modelos. Por último se proporcionan unas conclusiones generales respecto a este estado del arte en la sección 2.4.

Aunque es difícil realizar una comparación entre los sistemas presentes en la literatura (ver sección 2.4), se ha tratado de describir algunos sistemas como ejemplo para proporcionar una idea general del estado del arte y las alternativas propuestas.

2.1. Bases de datos de habla emocional

La selección de la base de datos de trabajo merece una atención especial durante el proceso de planificación del sistema de identificación de emociones, ya que se trata del elemento que determinará qué emociones será capaz de detectar y bajo qué condiciones podrá ser utilizado el sistema.

Las características más importantes de una base de datos de habla emocional son:

- **La naturaleza de las emociones.** Dependiendo de si las emociones grabadas son naturales o actuadas el sistema tendrá mayor o menor capacidad de identificar emociones en el habla espontánea.
- **El número y tipo de emociones.** Esto influye directamente en las emociones que el sistema será capaz de reconocer. Un mayor número de emociones a identificar también provoca una mayor probabilidad de confusión entre ellas.
- **El número de locutores grabados.** A mayor número de locutores, mayor será la capacidad del sistema para generalizar y detectar emociones en la voz de personas que no forman parte de la base de datos de entrenamiento.

- **La calidad de las grabaciones.** Una base de datos grabada en un entorno acondicionado, libre de ruido, facilita en gran medida la caracterización de las emociones. Sin embargo, hace que el sistema tenga más problemas si posteriormente es utilizado en condiciones reales de ruido ambiente.

Existen diferentes bases de datos disponibles para el desarrollo de sistemas de habla emocional, cada una con sus características particulares en función del objetivo para el que ha sido grabada (reconocimiento automático del habla (RAH), conversión de texto a habla (CTH), análisis de habla emocional). Esta sección analiza las características más importantes de las principales bases de datos disponibles. En la Tabla 2.1 al final de la sección puede encontrarse un resumen de estas características. No se pretende realizar una revisión exhaustiva de todas las bases de datos de habla emocional existentes, tan sólo de las más significativas para permitir comparar las diferentes características de las mismas. Cowie *et al.* (2005) y Ververidis y Kotropoulos (2006) presentan una revisión más completa con estas y otras bases de datos.

2.1.1. Naturaleza de las emociones

El mayor problema a la hora de grabar una base de datos emocional es conseguir que una persona muestre una emoción para poder grabarla. Suelen utilizarse cuatro técnicas diferentes para lograr que las emociones afloren durante la grabación de la base de datos (Campbell, 2000).

Emociones actuadas

Se trata de la técnica más sencilla, en donde los locutores leen un texto predefinido fingiendo la emoción indicada. Este tipo de grabaciones puede hacerse en una sala acondicionada, con lo que se obtiene un control total sobre las condiciones de grabación y el contenido de los textos. Por el contrario, proporciona poca naturalidad en la expresión de las emociones, generalmente debido a la sobreactuación. Puesto que el locutor no tiene ninguna otra forma de expresar la emoción solicitada que la propia voz, no puede apoyarse ni en gestos ni en el contenido del texto, tiende a exagerar para que las emociones sean identificables (Batliner *et al.*, 2000).

Esta técnica se utiliza sobre todo para grabar bases de datos orientadas a sistemas de CTH emocionada, donde obtener grabaciones de alta calidad es en general muy importante. Además, debido a la sobreactuación, estas bases de datos recogen expresiones emocionales muy identificables, adecuadas para desarrollar un sistema de CTH emocionada, donde lo importante es que el oyente reconozca la emoción pretendida.

También es una técnica muy utilizada para hacer análisis básico de la señal de voz emocional. Al tener un total control sobre los textos leídos se puede grabar el mismo texto en cada emoción, con lo que se facilita la comparación directa de las características en los diferentes estilos.

Emociones estimuladas

En este caso se hace leer al locutor un texto con un alto contenido emocional, de forma que esa emoción se refleje en la voz. Se espera que con un texto altamente emotivo las emociones se reflejen más fácilmente en la voz sin necesidad de sobreactuar, con lo que se obtiene una mayor naturalidad. Esta técnica permite mantener un alto control sobre las condiciones de grabación y el contenido de los textos, aunque se pierde la capacidad de grabar textos paralelos.

Emociones evocadas

Consiste en hacer que el locutor se ponga en una situación emotiva y la describa. Por ejemplo, haciéndole recordar momentos emotivos de su vida, o mostrándole una serie de imágenes con alto contenido emocional y pidiéndole que las comente.

Las grabaciones obtenidas mediante esta técnica son mucho más naturales, ya que no se pide al locutor que sus emociones sean reconocibles, y por tanto, no necesita sobreactuar. Por el contrario, se pierde la capacidad de controlar el contenido de las grabaciones, aunque todavía se puede utilizar una sala especial y controlar las condiciones de grabación.

Emociones reales

Las bases de datos de emociones reales suelen obtenerse a partir de programas de televisión con alto contenido emotivo (debates, programas de telerrealidad, etc.), a través sistemas de mago de Oz (WOZ, *Wizard of OZ*), o mediante entrevistas y debates preparados en los que el entrevistador o moderador trata de forzar discusiones de carga emotiva. Contienen emociones totalmente naturales, sin embargo, en este caso no se tiene ningún control sobre el contenido ni las condiciones de grabación.

Siendo un procedimiento totalmente incontrolado, en el que los locutores sienten las emociones de forma espontánea, es muy complicado lograr ejemplos de ciertas categorías. Generalmente suelen obtenerse ejemplos de habla neutra (la mayoría), ira/agresividad (debates, WOZ), tristeza (telerrealidad) y frustración (WOZ).

Esta técnica presenta además la dificultad añadida de tener que etiquetar manualmente las emociones grabadas, ya que no hay ninguna referencia sobre qué emoción está sintiendo el locutor en cada momento. Este etiquetado manual puede

ser muy difícil, puesto que las emociones humanas casi nunca son puras, y se mezclan muchas de ellas a la vez. Por lo tanto, no es sencillo seleccionar una única etiqueta para identificar el estado emocional del locutor en cada momento. Generalmente se necesita analizar el resultado de varios etiquetadores y buscar un consenso entre todos para asignar las etiquetas definitivas a las grabaciones.

Todas estas técnicas tratan de buscar un punto de trabajo óptimo entre la naturalidad de las emociones y el control sobre las grabaciones. En un extremo están las emociones actuadas, con un control total sobre el contenido y las condiciones de grabación pero con poca naturalidad debido a la sobreactuación. En el otro extremo están las emociones reales, totalmente naturales, pero sin ningún control sobre las grabaciones. La Figura 2.2 muestra de forma gráfica el compromiso de cada una de las técnicas.

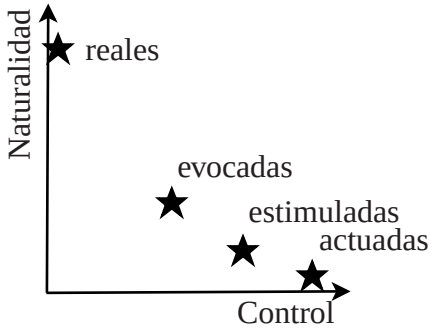


FIGURA 2.2: *Diagrama representando la naturalidad de la emoción frente al control sobre las grabaciones.*

En general las bases de datos actuadas proporcionan emociones arquetípicas muy intensas y totalmente separadas entre sí. Esto hace que sean más fáciles de distinguir una de otra. Sin embargo las emociones humanas rara vez se dan de forma aislada y con total intensidad, sino que generalmente lo que se siente es una conjunto de emociones combinadas. Esta mezcla hace que sean difíciles de distinguir entre sí en muchos casos. Por tanto, las bases de datos de emociones naturales proporcionan un marco realista para el desarrollo de sistemas de reconocimiento automático de emociones (RAE), donde el objetivo final es detectar el estado emocional de un locutor hablando de forma natural.

También es cierto que el ser humano muchas veces no experimenta las emociones con total intensidad, con lo que es aún más complicado determinar si una emoción está presente o no. A este respecto es interesante la base de datos de Iriondo *et al.* (2000), donde se recogen las emociones en tres intensidades diferentes, lo que puede facilitar un estudio de las características del habla emocionada

en función de la intensidad.

2.1.2. Número de emociones

Existe una corriente muy extendida en el campo de la psicología que separa las emociones entre básicas y secundarias (Cowie, 2000). Según esta corriente, las emociones básicas son aquellas que son universalmente conocidas en todas las culturas, presentan unas características muy diferentes entre sí, y son las únicas que pueden darse con total intensidad. Por su parte las emociones secundarias nunca son tan intensas y se presentan con características menos diferenciadas. Aunque no se ha llegado a un acuerdo sobre cuáles son las emociones básicas, por regla general se acepta que son las incluídas en el conjunto conocido como *the big six*: miedo, ira, alegría, tristeza, sorpresa y asco.

La mayoría de las bases de datos recogen todas o algunas de estas emociones básicas, aunque a veces se sustituye alguna de ellas por otra más oportuna para los objetivos para los que fue grabada. Es el caso de la base de datos *Berlin* (Burkhardt *et al.*, 2005), que sustituye el asco por el aburrimiento, o de la *Mozziconacci* (Mozziconacci y Hermes, 2000) que cambia asco y sorpresa por aburrimiento e indignación. Una de las emociones de la que más se prescinde es la sorpresa, tal y como se hace en la propia *Mozziconacci*, en *Magdeburger* (Wendt y Scheich, 2002), *Ruslana* (Makarova y Petrushin, 2002), y en las bases de datos de Amir *et al.* (2000), McGilloway *et al.* (2000) y Petrushin (2000). También el asco parece ser evitado muchas veces, ya que no está presente en *Berlin* (Burkhardt *et al.*, 2005), *DES* (Engberg y Hansen, 1996), *SES* (Montero *et al.*, 1999), *Mozziconacci* (Mozziconacci y Hermes, 2000) ni en las bases de datos de McGilloway *et al.* (2000) y Petrushin (2000). Tal vez la razón sea que el asco es una emoción muy difícil de detectar a través de la voz. Scherer (2003) explica esta dificultad en términos evolutivos, indicando que el asco se ha transmitido más eficientemente a través del rostro que la voz. Alternativamente, Banse y Scherer (1996) sugieren que la expresión natural del asco en la voz se refleja sobre todo mediante interjecciones cortas (por ejemplo, *ugh!*, *puaj!*) en lugar de mediante frases completas con características acústicas concretas.

Cuando la base de datos se graba para una aplicación concreta suele ocurrir que el número de emociones reflejadas se reduce al mínimo, recogiendo sólo aquellas que interesan. Es el caso de las bases de datos *SUSAS* (Hansen y Bou-Ghazale, 1997) y *Drivers* (Fernandez y Picard, 2000), ambas enfocadas a desarrollar sistemas de RAH en condiciones reales de ruido y estrés intenso. Estas bases de datos recogen básicamente voz estresada a volumen alto y con un pronunciado efecto Lombard.

Tampoco suele ser muy variado el contenido de las grabaciones logradas con el método de WOZ. En estos casos suele recogerse sobre todo enfado y frustración,

cuando la supuesta máquina no obedece a los comandos indicados, tal y como se refleja en *SmartKom* (Schiel *et al.*, 2002).

Por último, hay unos pocos grupos de investigación que han tratado de abrir el abanico de emociones grabadas hasta superar la decena. Buenos ejemplos son las bases de datos *GVEESS* (Banse y Scherer, 1996), *Belfast* (Douglas-Cowie *et al.*, 2000) y *EPST* (Liberman *et al.*, 1999). Además de las emociones consideradas básicas, estas bases de datos incluyen, entre otras, ira contenida, desesperación, desprecio, júbilo, interés, vergüenza, orgullo y pánico.

2.1.3. Número de locutores

El número de locutores que debe tener una base de datos de habla emocional depende de la aplicación que se vaya a dar a la misma. En el caso de querer desarrollar un sistema RAE, es conveniente disponer de muchos locutores para que sea capaz de aprender la variabilidad inter-locutor, y así poder reconocer las emociones expresadas por personas que no forman parte del conjunto de entrenamiento. Por el contrario, el desarrollo de sistemas de CTH emocionada puede realizarse con bases de datos de un único locutor, sobre todo si se trata de síntesis concatenativa.

Por tanto es lógico que el número de locutores varíe mucho en las diferentes bases de datos disponibles. Las dedicadas a sistemas CTH o a análisis general del habla son las que menos locutores contienen, no superando en muchos casos los cuatro locutores (*DES* (Engberg y Hansen, 1996), *Interface* (Hozjan *et al.*, 2002), *Magdeburger* (Wendt y Scheich, 2002), *Mozziconacci* (Mozziconacci y Hermes, 2000), *AhoEmo1* (Navas *et al.*, 2004b), *AhoEmo2* (Saratxaga *et al.*, 2006)).

En cuanto a las bases de datos dedicadas al desarrollo de sistemas RAE, contienen un número más elevado de locutores para recoger la variabilidad inter-locutor. La mayoría recogen más de 20 locutores, como *Belfast* (Douglas-Cowie *et al.*, 2000), *AIBO* (Batliner *et al.*, 2006), *SmartKom* (Schiel *et al.*, 2002), *Baggage loss* (Scherer y Ceschi, 2000) o las bases de datos de Amir *et al.* (2000), McGilloway *et al.* (2000) y Petrushin (2000). Sin embargo algunas se quedan por debajo, como *Berlin* (Burkhardt *et al.*, 2005), *GVEESS* (Banse y Scherer, 1996) y *ESMBS* (Nwe *et al.*, 2003).

2.1.4. Conclusiones del análisis de las bases de datos

Las principales características de las bases de datos de habla emocional vienen condicionadas sobre todo por el objetivo para el que fueron grabadas. Si este objetivo era el desarrollo de sistemas CTH, suelen contener muy pocos locutores (a veces tan sólo uno) y emociones actuadas, muchas veces notablemente exageradas. Aunque en ocasiones este tipo de bases de datos han sido utilizadas para el

diseño de experimentos de identificación automática de emociones, son sin duda alguna las menos adecuadas para ello, ya que los resultados obtenidos son muy poco realistas. Debido a la falta de locutores no es posible realizar experimentos independientes de locutor, por lo que el resultado no es extrapolable a situaciones reales. Además, al tratarse de emociones en su mayoría sobreactuadas, estos resultados tampoco son aplicables a la vida real, donde las manifestaciones emocionales son mucho más sutiles.

Idealmente los sistemas RAE deberían desarrollarse sobre bases de datos de emociones reales, no sobreactuadas. Sin embargo este tipo de bases de datos presenta como inconveniente un número generalmente reducido de emociones. Debido a la metodología utilizada para su grabación (debates, telerrealidad, WOZ) sólo se consigue recoger un escaso número de expresiones, generalmente negativas (enfado, tristeza, desesperación, frustración). Como contrapartida, suele tratarse de bases de datos con un elevado número de locutores, lo que sumado a la naturalidad de las emociones proporciona resultados muy realistas. Son, por tanto, muy adecuadas para el diseño de sistemas de identificación especializados en esas emociones negativas, como por ejemplo, sistemas automáticos de atención al cliente. Para sistemas más generalistas, capaces de interpretar otras expresiones emocionales, sería necesario escoger bases de datos menos realistas que amplían el abanico de emociones.

Aquellas bases de datos que tengan muy pocos locutores deberían ser descartadas para desarrollo de sistemas RAE, debido a la imposibilidad de diseñar experimentos independientes de locutor. Esta independencia de locutor es la única forma de extrapolar los resultados de los experimentos a situaciones reales en las que el sistema debe identificar la emoción de un locutor desconocido. A este respecto se puede considerar que una base de datos con menos de 8–10 locutores es inapropiada para los sistemas RAE.

TABLA 2.1: *Resumen de algunas bases de datos de habla emocional.*

Base de datos	#Emo	#Loc	Contenido	Tipo	Idioma	Objetivo	Observaciones
DES (Engberg y Hansen, 1996)	5	4	*	Actuadas	Danés	CTH	* Se grabaron 2 palabras, 9 frases y 2 párrafos por locutor y emoción.
GVEESS (Banse y Scherer, 1996)	14	12	280 frases	Estimuladas	Ficticio	RAE, Análisis	Inicialmente 1344 frases que se evaluaron en naturalidad e identificación, eliminando las que tenían una puntuación inferior al 40% en naturalidad o del 66% de tasa de identificación. Las frases son un conjunto de fonemas indo-europeos, pronunciables pero sin sentido.
SUSAS (Hansen y Bou-Ghazale, 1997)	Estrés*	32	16000 frases	Reales, Actuadas	Inglés	RAH	* Enfocada a aplicaciones de RAH bajo efectos de estrés y ruido. Básicamente contiene grabaciones bajo presión, ruido de fondo, efecto Lombard, etc.
EPST (Liberman *et al.*, 1999)	14	8	9 horas	Actuadas	Inglés	Análisis	
SES (Montero *et al.*, 1999)	5	1	*	Actuadas	Castellano	CTH	* 3 párrafos cortos, 15 frases y 30 palabras aisladas repetidas en cada emoción.
Lost luggage (Scherer y Ceschi, 2000)	Varias*	112	10-20 min/loc	Reales	Inglés, Francés	RAE	* Grabadas en el mostrador de maletas perdidas de un aeropuerto. Contiene principalmente irritación/enfado, resignación/tristeza, preocupación/estrés y buen humor.
Belfast (Douglas-Cowie *et al.*, 2000)	Varias	125	239 frases	Reales	Inglés	RAE	Grabaciones tomadas de entrevistas personales (30) y programas de televisión (209), expresando por tanto un gran abanico de emociones.
Driver (Fernandez y Picard, 2000)	2	4		Reales	Inglés	RAH	Similar a SUSAS, está enfocada a construir sistemas de RAH bajo situaciones de presión. Se grabó a personas que contestaban a problemas matemáticos mientras manejaban un simulador de conducción.

Continúa en la siguiente página

TABLA 2.1: *Continuación.*

Base de datos	Emo	#Loc	Contenido	Tipo	Idioma	Objetivo	Observaciones
Amir (Amir *et al.*, 2000)	5	31		Evocadas	Hebreo	RAE	Habla espontanea de recuerdos emotivos para el locutor.
Iriondo (Iriondo *et al.*, 2000)	7	8	2x3 textos/loc/emo*	Actuadas	Castellano	CTH	* Los textos se grabaron en 3 intensidades de emoción diferentes.
McGilloway (McGilloway *et al.*, 2000)	5	40	5 textos/loc/emo	Estimuladas	Inglés	RAE	
Mozziconacci (Mozziconacci y Hermes, 2000)	7	3	5 frases/loc/emo	Estimuladas	Holandés	Análisis	
Petrushin (Petrushin, 2000)	5	30	4 frases/loc/emo	Actuadas	Inglés	RAE	
Interface (Hozjan *et al.*, 2002)	7	2/idioma	175-190 frases/loc/emo	Actuadas	Inglés, Castellano, Francés, Esloveno	CTH, Análisis	Cooperación entre 4 sedes, cada una grabando en un idioma. Cada sede debía grabar como mínimo 6 emociones definidas, aunque luego algunas grabaron variaciones del neutro (rápido, alto, lento, etc.).
Magdeburger (Wendt y Scheich, 2002)	6	2	200 palabras/loc/emo	Actuadas	Alemán*	Análisis	* Palabras sin sentido construidas según las reglas de pronunciación alemanas.
Ruslana (Makarova y Petrushin, 2002)	6	61	10 frases/loc/emo	Actuadas	Ruso	RAE	
SmartKom (Schiel *et al.*, 2002)	Varias*	45	9 minutos/loc	Reales	Alemán	RAE	* Emociones reales tomadas mediante el método WOZ. Básicamente frustración, enfado y neutro.

Continúa en la siguiente página

TABLA 2.1: *Continuación.*

Base de datos	Emo	#Loc	Contenido	Tipo	Idioma	Objetivo	Observaciones
BabyEars (Slaney y McRoberts, 2003)	3	12	30-50 frases/loc	Reales	Inglés	RAE	Grabaciones de padres y madres interactuando con sus hijos de 10-18 meses.
Groningen (Choukri, 2003)	*	238	20 horas	Actuadas	Holandés	RAH	* Los textos leídos contienen muchas citas de caracter emotivo para provocar habla emocionada.
ESMBS (Nwe *et al.*, 2003)	6	12	10 frases/loc/emo	Actuadas	Birmano, Mandarín	RAE	6 locutores por cada idioma.
AhoEmo1 (Navas *et al.*, 2004b)	7	1	170 grabaciones/emoción*	Actuadas	Euskera	CTH, análisis	* Contiene frases largas y palabras aisladas, tanto semánticamente neutras como relacionadas con la emoción.
Berlin (Burkhardt *et al.*, 2005)	7	10	535 frases semánticamente neutras	Actuadas	Alemán	RAE	Las frases fueron evaluadas para garantizar su naturalidad y reconocibilidad.
AIBO (Batliner *et al.*, 2006)	Varias*	51	9 horas	Reales	Alemán	RAE	* Grabaciones de niños dirigiendo al robot *AIBO* mediante comandos de voz. Muchas emociones registradas, pero como algunas están muy poco representadas, suelen agruparse en 4 o 5 categorías.
AhoEmo2 (Saratxaga *et al.*, 2006)	7	2	702 frases/emoción	Actuadas	Euskera	CTH, análisis	

2.2. Parámetros para la identificación de emociones en el habla

El tipo de parametrización utilizada tiene una gran influencia sobre el comportamiento de un sistema de identificación de emociones. Los parámetros que se extraen de la señal de voz determinan en gran medida la capacidad del sistema para distinguir las emociones y para adaptarse a voces de locutores desconocidos. Una parametrización adecuada para la identificación de emociones debe tener:

- Una elevada **varianza inter–emoción**. Si los parámetros extraídos a partir de señales correspondientes a distintas emociones son muy diferentes entre sí, se facilita la discriminación entre estas emociones.
- Una reducida **varianza intra–emoción**. Los parámetros calculados a partir de señales de una misma emoción deben ser muy similares entre sí, de forma que se reduzca la confusión entre las diferentes emociones. Esto incluye que los parámetros sean estables frente a cambios de locutor, de forma que tengan valores similares independientemente de la identidad de la persona que habla. En caso contrario el sistema no podrá identificar correctamente las emociones de locutores para los que no haya sido entrenado.

Uno de los mayores problemas a la hora de decidir los parámetros a utilizar es que no hay una base teórica sólida que relacione el estado emocional del locutor con las características de su voz (Scherer, 2003). La mayoría de los trabajos llevados a cabo en este campo se basan en parámetros obtenidos a partir de la comparación directa de señales de voz con diferentes emociones. Esta técnica permite estimar las diferencias acústicas entre las señales, y por tanto, determinar qué parámetros pueden servir para identificar esas emociones. Sin embargo, no disponer de la base teórica para relacionar la emoción con estas características acústicas supone una gran desventaja. Debido a la complejidad de la señal de voz, nunca se puede estar seguro de que las diferencias encontradas mediante comparación directa sean debidas al estado emocional del locutor o a otros factores como el entorno, ruido, diferencias en el equipamiento o identidad del locutor. También es muy posible que otras características diferenciadoras no hayan sido descubiertas al estar ocultas bajo la complejidad de la señal.

La falta de esta teoría acústica de las emociones proviene de la propia teoría emocional. La psicología no ha llegado a un acuerdo sobre qué es una emoción, mucho menos, cómo llega a afectar a nuestra mente y cuerpo (Cornelius, 2000). Sólo unos pocos autores han intentado desarrollar una teoría que describa los cambios fisiológicos provocados por el estado emocional (ver por ejemplo los trabajos de Scherer (2000) y Ekman (1992)). Estos trabajos toman como referencia la teoría Darwiniana (Darwin, 1872), que afirma que las emociones son producto de una

necesidad evolutiva. Según esta teoría, los cambios fisiológicos provocados por un cierto estado emocional responden a la necesidad de preparar el cuerpo ante los factores que provocan dicha emoción. Por ejemplo, sentimos miedo para prepararnos ante un peligro. Para ello el cuerpo responde incrementando la segregación de adrenalina, lo que provoca un aumento de la frecuencia cardíaca, la presión sanguínea y la tensión muscular. También se abre la boca para aumentar el flujo de aire y se corta la salivación. Todo ello hace que el cuerpo esté dispuesto para enfrentarse a un peligro, bien sea huyendo o contraatacando. Todos estos cambios tienen un efecto sobre la fisiología del aparato fonador y esto se refleja en las características de la voz (Nwe *et al.*, 2003; Williams y Stevens, 1981). Al tensarse los músculos también lo hacen las cuerdas vocales y los musculos responsables de la respiración, provocando un incremento en el tono y en la intensidad del sonido. Al abrirse la boca y cortarse la salivación, el tracto vocal se seca, lo que junto a la tensión muscular hace que cambie la estructura del tracto vocal. Esto se refleja en una alteración del espectro de la señal. Además la tensión nerviosa provocada por la adrenalina hace que se hable más rápido y con pausas más cortas. Otras emociones, como la tristeza, tienen un efecto contrario, incrementando la salivación y relajando los músculos. Como consecuencia la entonación es más grave y la voz menos potente. En este caso no hay tensión nerviosa, sino relajación, con lo que el habla es más lenta y pausada.

En vista de los efectos descritos sobre el organismo, durante muchos años se han utilizado medidas de entonación, energía y velocidad del habla casi de forma exclusiva en sistemas automáticos de identificación de emociones. La utilización de este tipo de parámetros proporciona un patrón de aciertos y confusiones entre emociones muy característico y recurrente a lo largo de la literatura. Por ejemplo, estas medidas permiten separar fácilmente los estados emocionales de mucha tensión (enfado, alegría) de los de baja tensión (tristeza, aburrimiento), pero dentro de cada grupo la confusión es muy grande (Scherer, 2003). Sin embargo, el ser humano es capaz de distinguir el enfado de la alegría y la tristeza del aburrimiento con bastante acierto, por lo que se refuerza la idea de que hay otras características de la voz útiles para la discriminación de emociones. Algunos trabajos han utilizado medidas espectrales (Kim *et al.*, 2007; Vlasenko *et al.*, 2007) o de calidad de la voz (Steidl *et al.*, 2005), verificando la importancia de estos parámetros a la hora de distinguir emociones.

En esta sección se recogen las parametrizaciones más utilizadas en la literatura para la identificación de emociones en el habla. A modo de resumen la Tabla 2.2 al final del capítulo recoge los diferentes trabajos revisados y sus características principales.

2.2.1. Parámetros prosódicos

La prosodia se refiere a la estructura supra-segmental de la entonación, energía y ritmo del habla. La estructura prosódica viene en parte definida por el contenido semántico y lingüístico de la oración, en forma de acentos, patrones entonativos (por ejemplo, para caracterizar frases interrogativas) o pausas prosódicas para clarificar el mensaje (equivalentes a comas o puntos en lenguaje escrito). Sin embargo, la prosodia también es utilizada con frecuencia para transmitir información no verbal. Si a esto se añade que muchos de los efectos fisiológicos descritos anteriormente influyen directamente en las características prosódicas, no es de extrañar que los parámetros prosódicos sean los más utilizados a la hora de tratar de identificar emociones.

El estudio presentado por Banse y Scherer (1996) es muy representativo a este respecto. En él se analizan las características acústicas de grabaciones de voz emocionada, ajustando los valores de varios parámetros prosódicos mediante una regresión lineal. Se toman como variables independientes todos aquellos factores que pueden afectar al valor del parámetro: la identidad del locutor, su sexo, el texto de la frase, el entorno y por supuesto, la emoción. En los resultados se aprecia que la emoción explica el 55 % de la variación de la intensidad media y el 50 % de la variación de la media de F_0, haciendo evidente la influencia de la emoción sobre los parámetros prosódicos. No es el único trabajo que refleja esta influencia, pues ha sido puesta de manifiesto en numerosos estudios (Burkhardt y Sendlmeier, 2000; Devillers *et al.*, 2004; Hashizawa *et al.*, 2004; Huang y Akagi, 2008; Navas *et al.*, 2004a; Paeschke, 2004; Paeschke *et al.*, 1999; Scherer *et al.*, 1991; Schröder, 2003). Erickson (2005) presenta un resumen bastante completo de estos trabajos y sus conclusiones.

Las alternativas a la hora de definir los parámetros prosódicos son innumerables, y prácticamente existen tantas definiciones de parámetros prosódicos como trabajos publicados en este campo. La mayoría de las veces suelen darse en forma de estadísticos a largo plazo, calculados generalmente a lo largo de toda la frase. Entre los más habituales encontramos estadísticos sencillos como el valor medio, varianza, mínimo, máximo, rango o líneas de tendencia de las curvas de entonación y energía; o la duración media de las sílabas como estimación del ritmo. Aunque también es posible utilizar otros estadísticos más elaborados con la esperanza de que retengan mayor información emocional: estadísticos de mayor orden, o que caractericen una parte significativa de la curva, como la pendiente de entonación en la última sílaba.

En cuanto a los trabajos publicados en la literatura que utilizan parámetros prosódicos para la identificación de emociones, la tarea más sencilla consiste en determinar si una cierta emoción está presente o no. Esto reduce el problema a dos clases, estilo neutro o emocionado, con lo que se reduce también la confusión del

clasificador. Esta tarea, aunque sencilla, tiene mucha aplicación en la detección de clientes enfadados o frustrados en sistemas de atención automática. Lee *et al.* (2001) desarrollaron un sistema de estas características parametrizando cada frase con un único vector de estadísticos prosódicos (6 parámetros para voces masculinas, 7 para voces femeninas). Con ello obtuvieron un 80 % de acierto en la detección del enfado sobre una base de datos de habla natural. Por desgracia, el resultado no es del todo generalizable, puesto que las pruebas se realizaron en un sistema dependiente de locutor. Yacoub *et al.* (2003) presentan un sistema similar, también enfocado a la detección de locutores enfadados. Sin embargo, esta vez las pruebas se realizan de forma independiente de locutor, obteniendo un 94 % de aciertos. El incremento en la tasa de acierto se explica en parte por la utilización de un mayor número de parámetros (hasta 37), y sobre todo, por realizar los experimentos sobre una base de datos de habla actuada.

Cuando el número de emociones consideradas se incrementa, los resultados obtenidos son mucho más discretos. Ververidis y Kotropoulos (2005) y McGilloway *et al.* (2000) informan de una precisión alrededor del 55 % con 5 emociones y una arquitectura dependiente de locutor, utilizando 7 y 32 parámetros respectivamente. Petrushin (2000) alcanza un 65 % de acierto con 14 parámetros, también con 5 emociones y una arquitectura dependiente de locutor. Seppänen *et al.* (2003) reducen el número de emociones a 4, con lo que logran alcanzar el 75 % de precisión con 10 parámetros. En este mismo trabajo, al implementar una arquitectura independiente de locutor, la precisión cae hasta el 60 %.

Hozjan y Kacic (2003) utilizan una gran cantidad de estadísticos (hasta 144) calculados a partir de las características prosódicas de la señal, en lo que ellos llaman el Gran Conjunto de Parámetros Estadísticos (LSSF, *Long Set of Statistical Features*). Con ello tratan de recoger la mayor cantidad posible de información acústica a la hora de discriminar las emociones. Para probar esta parametrización utilizan la base de datos *InterFace*, que contiene 7 emociones diferentes, en una arquitectura de locutor único (el entrenamiento y las pruebas se realizan con un único locutor). Bajo estas circunstancias obtienen una precisión que varía entre el 60 % y el 90 % en función del idioma y locutor. Sin embargo, todo hace suponer que en caso de probarse en una arquitectura independiente de locutor los resultados serían mucho más discretos.

Como ya se ha indicado, la prosodia no depende únicamente de la emoción. Se trata de una característica que transporta una gran cantidad de información lingüística en forma de acentos, patrones entonativos y pausas. Esto quiere decir que la prosodia de una frase depende en gran medida del contenido semántico de la misma. A la hora de identificar las emociones, esta información lingüística supone ruido que puede ocultar las características emocionales. Jiang y Cai (2004) han tratado de eliminar la información lingüística de la prosodia de la señal como paso previo a realizar la detección de emociones. Para ello han usado un sistema

de predicción prosódica similar al utilizado en sistemas CTH, que proporciona una estimación de la prosodia en estilo neutro para el texto de entrada. Puesto que se trata de una estimación para el estilo neutro, se supone que toda la información prosódica existente es debida al contenido de la frase. De esta forma son capaces de eliminar la información lingüística de la prosodia de la frase original y utilizar el residuo para la clasificación de la emoción. Con este método y utilizando sólo 8 parámetros prosódicos consiguen una precisión de 93,7 % en una base de datos de 6 emociones. Aunque se trata de un sistema de locutor único, ya que todo el sistema (incluido el predictor prosódico) está entrenado para un único locutor, el resultado es bastante significativo teniendo en cuenta el reducido número de parámetros utilizados. Por desgracia, no se aporta una comparación con la precisión del sistema sin utilizar el predictor.

Existen muchos más trabajos que utilizan parámetros prosódicos para la identificación, de hecho la mayoría lo hacen (ver Tabla 2.2). Estos estudios utilizan la prosodia en combinación con otros parámetros, por lo que serán comentados en la sección correspondiente a esos otros parámetros, para facilitar la comparación y evitar repeticiones.

2.2.2. Parámetros espectrales

Los efectos fisiológicos derivados del estado emocional afectan también al tracto vocal, provocando variaciones en las características espectrales de la voz. Aunque se trata de variaciones menos pronunciadas que en el caso de la prosodia, pueden ser utilizadas para la identificación de las emociones. Efectivamente, el ya citado estudio de Banse y Scherer (1996) muestra que la emoción explica parte de la variación de la forma espectral de la señal. Sin embargo, en este caso una parte significativa de la variación está asociada a la identidad del locutor. Lo cual es lógico, pues los parámetros espectrales tales como MFCC (coeficientes cepstrales en escala mel, *Mel Frquency Cepstral Coefficients*) o LPCC (coeficientes cepstrales de predicción lineal, *Linear Prediction Cepstral Coefficients*) son muy utilizados en sistemas de identificación de locutor (Bimbot *et al.*, 2004; Campbell, 1997; Lu y Dang, 2008; Orman y Arslan, 2001; Reynolds y Rose, 1995).

La importancia de los parámetros espectrales también se demuestra de forma empírica a través de los numerosos trabajos de identificación de emociones que hacen uso de ellos. Casale *et al.* (2007) por ejemplo han desarrollado un sistema para la detección de estrés utilizando la base de datos *SUSAS*. Utilizando únicamente parámetros MFCC y modelos ocultos de Markov (HMM, *Hidden Markov Models*) consiguen un 76,8 % de acierto. Puede compararse este resultado con los obtenidos mediante parámetros prosódicos en trabajos similares anteriormente citados, como los de Lee *et al.* (2001) y Yacoub *et al.* (2003), que alcanzaban un 80 % y 94 % de precisión respectivamente detectando el enfado. El resultado de

Casale *et al.* es ligeramente inferior, pero verifica que ciertamente los parámetros espectrales tienen algo que aportar a la identificación de emociones. Estos resultados quedan confirmados por los trabajos de otros autores que también utilizan exclusivamente parámetros espectrales para la detección de emociones. Es el caso de Truong y van Leeuwen (2007), que consiguen identificar 7 emociones con un 75 % de precisión utilizando parámetros rasta-PLP; o el de Nwe *et al.* (2003), que discriminan entre 6 emociones utilizando LFPC (coeficientes de potencia en escala logarítmica, *Log-Frequency Power Coefficients*) con una tasa de acierto entre el 75 % y el 80 %.

A menudo se trata de combinar los parámetros prosódicos y espectrales para obtener mejores resultados en la identificación. Teniendo en cuenta que la prosodia es generada casi exclusivamente por la actividad de las cuerdas vocales, y que la forma espectral tiene mayor influencia del tracto vocal, es razonable suponer que ambas características están poco correladas y que transmiten diferente información acerca de la emoción. Por tanto, es de esperar que la combinación de ambas fuentes de información sea provechosa. El problema de combinar la información prosódica y espectral es la diferente naturaleza de ambas. Mientras que la prosodia proporciona información a largo plazo, la forma espectral se ha procesado tradicionalmente a nivel segmental, generando un vector de parámetros por cada trama de 20-30 ms. Combinar parámetros de diferente resolución temporal de forma directa no es posible, por lo que diferentes autores aplican diferentes mecanismos de fusión. Algunos prefieren utilizar la fusión a nivel de parámetros o fusión temprana (*early fusion*) mientras que otros utilizan fusión a nivel de clasificador o fusión tardía (*late fusion*).

Una de las soluciones más simple y utilizada es calcular estadísticos a largo plazo para los parámetros espectrales y concatenar directamente los parámetros prosódicos y espectrales. Con esta técnica Grimm *et al.* (2007) tratan de separar 4 emociones con un total de 46 parámetros, alcanzando un 84 % de precisión en una arquitectura dependiente de locutor. El resultado baja hasta el 66,9 % cuando es independiente de locutor. Este resultado es similar al obtenido por Pierre-Yves (2003), también para una base de datos de 4 emociones y una arquitectura dependiente de locutor, aunque en este caso utiliza hasta 200 parámetros prosódicos y espectrales. En este trabajo se comparan varios clasificadores diferentes, obteniendo entre un 80 % y un 95 % de acierto en función del clasificador.

Como es lógico, los resultados descienden un poco si se intenta identificar un mayor número de emociones. Chichosz y Slot (2007) combinaron estadísticos de prosodia y energía por bandas de frecuencia para identificar las emociones de la base de datos *Berlin* (7 emociones). Con esto lograron un 74 % en una arquitectura dependiente de locutor y un 72 % en una independiente de locutor. Resultados similares consiguieron Shami y Verhelst (2007) también sobre la base de datos *Berlin*: 76 % de acierto en una arquitectura dependiente de locutor usando

102 parámetros. Yendo más allá, Vogt y André (2006) parten de un vector de 1280 estadísticos de prosodia y MFCC y aplican un procedimiento de selección de parámetros para retener los 50 más discriminantes. Con esta técnica logran un 81 % de acierto en una arquitectura independiente de locutor sobre las 7 emociones de la base de datos *Berlin*. Este incremento en la tasa de acierto posiblemente sea debida a haber partido de un número tan elevado de parámetros.

En lugar de utilizar los tradicionales parámetros MFCC o la energía por bandas de frecuencia, hay quien prefiere utilizar parámetros de articulación. Es la aproximación utilizada por Morrison *et al.* (2007), que combinan 38 estadísticos de prosodia y formantes. Con estos parámetros logran un 72 % de acierto en una base de datos de 6 emociones actuadas y una arquitectura dependiente de locutor. El mismo artículo describe la aplicación del sistema sobre una base de datos natural con sólo dos emociones, ira y neutro, obteniendo un 79 % de precisión.

Otra alternativa para la fusión temprana de la información prosódica y espectral es hacerlo al contrario. En lugar de concatenar estadísticos de prosodia y espectro para crear un único vector por frase, se pueden concatenar los parámetros espectrales, calculados por cada trama, con las primitivas de la prosodia (las muestras de F_0 y energía), calculadas también trama a trama. De esta forma, en lugar de tener un único vector por frase se obtiene un vector por cada trama. Kwon *et al.* (2003) comparan las dos técnicas en dos bases de datos, *SUSAS* (4 emociones) y *AIBO* (5 emociones), utilizando parámetros extraídos de MFCC, energía por bandas de frecuencia, formantes, intensidad y F_0. En el caso de los parámetros a largo plazo usan una máquina de vectores soporte (SVM, *Support Vector Machine*) como clasificador, logrando un 42 % de precisión en *AIBO* y un 67 % en *SUSAS*. También utilizan la base de datos *SUSAS* para tratar de distinguir únicamente entre estrés y neutro, con un 91 % de acierto. En el caso de los parámetros a nivel de segmento utilizan HMM y consiguen un 41 % en *AIBO*, un 70 % en *SUSAS* y un 96 % de detección de voz estresada. Con unos resultados tan similares es difícil saber cuál de las dos aproximaciones es más adecuada. Los experimentos con parámetros a corto plazo parecen ser mejores sólo cuando hay pocas emociones a distinguir. Sin embargo, puede que la diferencia sea debida a utilizar un clasificador diferente y no a la parametrización propiamente dicha. Por ejemplo, Barra-Chicote *et al.* (2009) obtienen un 67 % de precisión sobre las 5 emociones de la base de datos *AIBO* utilizando parámetros espectrales y primitivas de prosodia a nivel de segmento con un clasificador SVM.

Por último, hay autores que prefieren mantener cada tipo de parámetro en su ámbito, utilizando los datos prosódicos a largo plazo y los espectrales a corto plazo, y aplicar una fusión tardía (Alkoot y Kittler, 1999; Kittler *et al.*, 1998; Ruta y Gabrys, 2000). En este caso, se desarrollan dos sistemas independientes, cada uno con una parametrización, y se combinan los resultados de la clasificación. Es la aproximación utilizada por Kim *et al.* (2007), que utilizan la regla de

la suma (Kittler *et al.*, 1998) para la fusión. De esta forma desarrollan un sistema para detectar voz enfadada, logrando entre un 83 % y un 95 % de acierto. También Vlasenko *et al.* (2007) utilizan esta técnica, alcanzando un 84 % de acierto tratando de separar voz estresada y neutra sobre la base de datos de habla natural *SUSAS*; y un 90 % usando las 7 emociones de la base de datos *Berlin*.

2.2.3. Parámetros de calidad de voz

La calidad de voz está formada por todas aquellas características de la voz que pueden derivarse a partir de la señal glotal, a excepción de la frecuencia fundamental (que es considerada dentro de los parámetros prosódicos). Como ya se ha indicado al inicio de la sección, las emociones tienen un gran efecto sobre las cuerdas vocales, y por tanto, es lógico que afecten a la calidad de voz: la tensión provocada por el miedo puede hacer que la voz suene entrecortada, mientras que la relajación del aburrimiento puede provocar una voz más susurrante. Esta relación entre emoción y calidad de voz queda reflejada en trabajos como los de Johnstone y Scherer (1999) y Gobl y Chasaide (2003). Sin embargo hay relativamente pocos autores que deciden utilizar este tipo de parámetros. Tal y como explican los propios Gobl y Chasaide, la razón fundamental es que obtener la señal glotal a partir de una grabación de voz es una tarea complicada que sólo puede hacerse con una precisión aceptable en regiones muy estables de la señal. Como resultado, es difícil estimar las características de calidad de voz con la precisión y estabilidad necesarias.

En cualquier caso, algunos investigadores han decidido utilizar la calidad de voz para caracterizar las emociones, con relativo éxito. Lugger y Yang (2007) consiguen alcanzar un 75 % de acierto discriminando entre 6 emociones actuadas en una arquitectura independiente de locutor, utilizando una combinación de parámetros prosódicos y de calidad de voz. Ofrecen además los resultados de la clasificación utilizando sólo la prosodia (67 %) y sólo la calidad de voz (61 %), certificando así que la contribución de esta última es significativa. Tato *et al.* (2002) presentan otro sistema implementado sobre la base de datos *AIBO*, capaz de distinguir entre 5 emociones con un 60 % de precisión utilizando una combinación de parámetros prosódicos y de calidad de voz. Aunque obtienen menor precisión que Lugger y Yang, hay que tener en cuenta que *AIBO* es una base de datos de habla espontánea. Cuando se compara este resultado con el obtenido por Kwon *et al.* (2003), sobre la misma base de datos *AIBO*, pero utilizando una combinación de parámetros prosódicos y espectrales a largo plazo (42 %), es cuando puede apreciarse el efecto de la calidad de voz.

2.2.4. Parámetros lingüísticos

Todos los parámetros analizados hasta ahora son de naturaleza acústica. Su ventaja es que se pueden obtener analizando las características acústicas de la señal, independientemente del contenido lingüístico de la frase. Sin embargo, este contenido lingüístico también ofrece ciertos indicios acerca del estado emocional del locutor, sobre todo en habla espontánea, ya que la repetición de ciertas palabras, interjecciones o eventos vocales (risas, suspiros, etc.) puede ser característica de sentimientos de ira, alegría o tristeza.

El mayor problema para utilizar parámetros derivados del contenido lingüístico es precisamente la necesidad de conocer este contenido, lo que sólo puede hacerse mediante un sistema de RAH. Los sistemas de RAH aumentan la complejidad de la arquitectura, y cometen errores que pueden dar lugar a una mala clasificación si se usan parámetros derivados de este reconocimiento. Además, este tipo de parametrizaciones son completamente dependientes del idioma considerado, y no pueden ser aplicadas a otras lenguas sin realizar primero un reentrenamiento del sistema. Sin embargo, si se aplican con cuidado parecen dar resultados positivos, sobre todo cuando los parámetros lingüísticos se combinan con parámetros acústicos. Müller *et al.* (2004) utilizan este tipo de parámetros junto con estadísticos a largo plazo de parámetros prosódicos sobre una base de datos de 7 emociones actuadas. Usando sólo parámetros semánticos alcanzan un 60 % de acierto, mientras que sólo con estadísticos prosódicos llegan al 74 %. Sin embargo, al combinar ambos tipos de información la tasa de acierto se incrementa hasta el 92 %. Un resultado similar (93 %) consiguen Schuller *et al.* (2005), también sobre 7 emociones actuadas en una arquitectura dependiente de locutor, usando parámetros semánticos y estadísticos de prosodia y espectro. Este mismo trabajo describe experimentos llevados a cabo en una arquitectura independiente de locutor, donde la tasa de acierto se reduce hasta el 71 %. En una base de datos con sólo dos emociones, enfocada a la detección de usuarios enfadados en sistemas de atención automática, López-Cozar *et al.* (2008) alcanzan el 94 % de precisión combinando estadísticos prosódicos con parámetros acústicos a corto plazo y medidas derivadas del léxico de las frases y los patrones de diálogo.

Por otro lado, este tipo de parámetros sólo tienen sentido al tratar de identificar emociones en el habla espontánea, donde el contenido lingüístico de las grabaciones varía. Por desgracia la mayoría de los trabajos se desarrollan sobre bases de datos de habla actuada leída, a menudo usando los mismos textos para todas las emociones, por lo que son pocos los autores que deciden utilizar este tipo de características.

2.2.5. Conclusiones del análisis de parámetros

Un repaso de la literatura muestra la falta de un consenso claro sobre cuáles son los mejores parámetros a utilizar en la identificación automática de emociones, provocando que cada sistema desarrollado utilice su propio conjunto de parámetros. Aunque esta diversidad de parametrizaciones puede estar originada por la falta de una teoría sólida que relacione el estado emocional de un locutor con los cambios en las características de su voz, el problema se ha agravado por la ausencia de un estudio sistemático que analice la efectividad de cada tipo de parametrización bajo las mismas condiciones. Los trabajos descritos en la literatura son muy difíciles de comparar, debido a las diferentes condiciones en las que se han desarrollado: número y naturaleza de las emociones, dependencia del locutor, número de parámetros utilizado, sistemas de clasificación, etc. Por lo tanto, no es posible determinar si las diferencias presentes en las tasas de acierto son debidas a la utilización de un determinado tipo de parámetros o a otros factores. Incluso en los pocos trabajos en los que se proporcionan resultados con diferentes parámetros usando una misma base de datos, se utilizan diferentes clasificadores y número de parámetros con cada parametrización (Kim *et al.*, 2007; Luengo *et al.*, 2009a, 2005; Vlasenko *et al.*, 2007).

Es cierto que está ampliamente aceptado que los parámetros prosódicos proporcionan la mayor parte de la información emocional, y que la combinación de parametrizaciones de diferente naturaleza permite mejorar la precisión de la clasificación. Sin embargo, algunos de los resultados citados contradicen estas conclusiones, presentando mejores resultados al utilizar parámetros espectrales (Casale *et al.* (2007), 83 %; Truong y van Leeuwen (2007), 75 %) que prosódicos (McGilloway *et al.* (2000), 52 %; Ververidis y Kotropoulos (2005), 56 %; Petrushin (2000), 65 %; Seppänen *et al.* (2003), 75 %)[1].

Respecto a la combinación de parámetros, la fusión tardía parece tener cierta ventaja, ya que los trabajos revisados en los que se aplica esta técnica están entre los que obtienen mejores resultados. Así Kim *et al.* (2007) obtienen un 95 % de acierto entre dos emociones en una arquitectura independiente de locutor. Similarmente López-Cozar *et al.* (2008) alcanzan un 94 % también entre dos emociones y Müller *et al.* (2004) llegan hasta un 92 % en una base de datos de 7 emociones.

[1] Estos resultados se corresponden a experimentos bajo condiciones comparables, con entre 4 y 6 emociones actuadas y una arquitectura dependiente de locutor. La única excepción es Truong y van Leeuwen (2007) que trabaja sobre 7 emociones en una arquitectura independiente de locutor.

2.3. Clasificadores utilizados

Un clasificador tiene como objetivo decidir la clase a la que pertenece una muestra desconocida. Para ello utiliza una serie de muestras parametrizadas de entrenamiento y busca las fronteras entre las clases, particionando el espacio de parámetros. Durante la clasificación, comprueba la región del espacio en la que se sitúan los parámetros de la muestra desconocida y determina la clase correspondiente.

Por tanto, la elección del clasificador a utilizar viene condicionada en gran medida por la naturaleza de la parametrización y las características de la base de datos de entrenamiento: número de clases a diferenciar, número de muestras de entrenamiento, parámetros secuenciales en el tiempo (segmentales) o globales (supra-segmentales), etc. Cada clasificador tiene una capacidad diferente para adaptarse a estas características, por lo que se deberá seleccionar aquel que ofrezca las mejores condiciones en función de la base de datos y parametrización seleccionadas.

Existen muchos sistemas de clasificación desarrollados (Duda *et al.*, 2001), y casi todos han sido utilizados alguna vez para la identificación de emociones. En esta sección sólo se contemplan aquellos clasificadores más habitualmente utilizados.

2.3.1. El problema del sobreentrenamiento

La búsqueda de las fronteras entre clases se realiza mediante un criterio de minimización de error: la frontera óptima es aquella para la que el error de clasificación es mínimo. Con el objetivo de simplificar el problema, supongamos que sólo hay dos clases, con etiquetas $C = \{-1, +1\}$, de forma que a cada vector de parámetros $\mathbf{x} \in \mathbb{R}^N$ le corresponde una etiqueta $y \in C$. Sea $\hat{y} = f(\mathbf{x})$ la función de decisión utilizada para clasificar el vector $\mathbf{x}$. El error medio cometido es:

$$E = \int |f(\mathbf{x}) - y| P(\mathbf{x}, y)\, \mathrm{d}\mathbf{x}\, \mathrm{d}y \tag{2.1}$$

Sin embargo, la distribución conjunta de vectores y etiquetas $P(\mathbf{x}, y)$ es desconocida, y sólo se puede inferir a partir de los datos de entrenamiento $\tilde{\mathbf{x}}_i$ y sus etiquetas conocidas $\tilde{y}_i$. Por lo tanto, tan sólo es posible determinar una aproximación de este error, el llamado *error empírico* o error de entrenamiento. Suponiendo que haya L vectores de entrenamiento, este error es:

$$\hat{E} = \frac{1}{L} \sum_{i=1}^{L} |f(\tilde{\mathbf{x}}_i) - \tilde{y}_i| \tag{2.2}$$

El problema es que esta aproximación sólo es válida si se tiene un número suficientemente grande de muestras de entrenamiento. Con pocas muestras se corre el peligro de estimar incorrectamente la distribución de las clases y hallar una frontera poco adecuada para el problema. Por ejemplo, siempre es posible forzar un error empírico nulo, a costa de hallar una frontera compleja que clasifique cada muestra de entrenamiento correctamente (Figura 2.3). El problema es que cuando se trata de clasificar una nueva muestra no vista durante el entrenamiento, la probabilidad de que se cometa un error es muy grande. Se dice entonces que el sistema está *sobreentrenado*, es decir, está excesivamente particularizado para la distribución de las muestras de entrenamiento y no para la distribución original de las clases, por lo que generaliza mal para muestras desconocidas.

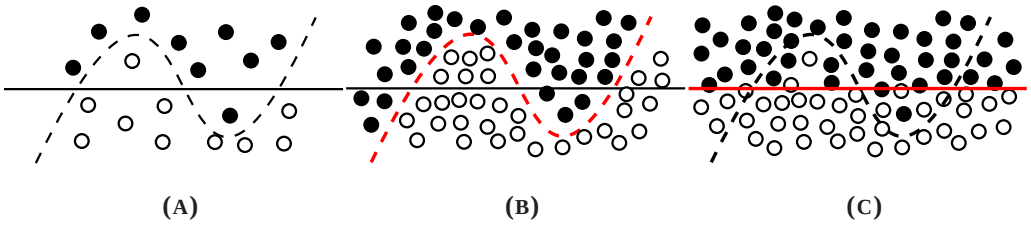


FIGURA 2.3: *Ejemplo gráfico del problema del sobreentrenamiento. Cuando se dispone de pocas muestras de entrenamiento (a) tanto la frontera simple (línea sólida) como la compleja (línea discontinua) pueden ser válidas. Sólo cuando se dispone de más muestras es posible determinar cuál de las dos era la correcta (b)- (c).*

Dicho de otra manera, el criterio de minimizar el error empírico no puede ser el único a tener en cuenta. Por suerte, hay métodos para paliar este problema. Por ejemplo, se puede buscar un compromiso entre el error empírico y la complejidad del sistema (Figura 2.3). Intuitivamente se entiende que una frontera sencilla que sea capaz de separar la mayoría de los datos de entrenamiento es más adecuada y estará menos sobreentrenada que una frontera compleja que consiga reducir el error empírico a cero. Los criterios AIC (*An Information Criterion*) (Akaike, 1974) y su versión mejorada BIC (*Bayesian Information Criterion*) (Schwarz, 1978) son ejemplos de métodos para alcanzar este compromiso.

Sin embargo, muchas veces es más sencillo buscar el equilibrio entre error y complejidad de forma empírica, mediante pruebas de validación cruzada. En este caso se divide la base de datos de entrenamiento aleatoriamente en M bloques, de forma que se utilizan $M-1$ bloques para entrenar el sistema, y el restante para realizar pruebas de validación. Este proceso se repite M veces, rotando cada vez el bloque de pruebas. Debido a que el bloque de pruebas no ha sido utilizado en el entrenamiento, los errores cometidos proporcionan una estimación de la ca-

pacidad de generalización del sistema para muestras no vistas, y por tanto, una estimación del error real. El método consiste en entrenar varios sistemas, modificando alguno de sus parámetros, y seleccionar aquel que proporcione menor error en las pruebas de validación.

2.3.2. Modelos de mezcla de gaussianas (GMM)

Los modelos de mezcla de gaussianas (GMM, *Gaussian Mixture Models*) tratan de estimar la función de densidad de probabilidad (fdp) de los parámetros pertenecientes a cada clase c mediante una suma ponderada de M distribuciones gaussianas (Paalanen *et al.*, 2006):

$$P_c(x) = \sum_{k=1}^{M} \omega_k^c \mathcal{N}(x;\ \mu_k^c,\ \Sigma_k^c) \tag{2.3}$$

siendo ω_k^c el peso de la componente k, con las siguientes condiciones:

$$\omega_k^c > 0 \qquad \sum_{k=1}^{M} \omega_k^c = 1 \tag{2.4}$$

El entrenamiento del modelo consiste en la estimación de los pesos ω_k^c y de los parámetros μ_k^c y Σ_k^c, y se realiza mediante el algoritmo EM (*Expectation Maximization*). La Figura 2.4 presenta como ejemplo un GMM en un espacio en dos dimensiones y cómo es capaz de aproximar una fdp concreta. Dado un número suficientemente alto de componentes, los GMM permiten aproximar cualquier tipo de distribución continua, independientemente de su forma.

La clasificación se basa en calcular las probabilidades $P(c|x)$ de que una muestra x pertenezca a cada una de las clases c, y seleccionar la más probable. Aplicando la regla de Bayes:

$$\hat{c} = \arg\max_c P(c|x) \ = \ \arg\max_c \frac{P(x|c)P(c)}{P(x)} \ = \ \arg\max_c P(x|c)P(c) \tag{2.5}$$

donde $P(c)$ es la probabilidad *a priori* de la clase c. En el caso de que la parametrización proporcione T vectores de parámetros por cada muestra a clasificar, se considera que cada uno de los vectores es independiente de los demás, con lo que la probabilidad conjunta se aproxima por:

$$P(x|c) = \prod_{t=1}^{T} P(x_t|c) \tag{2.6}$$

Como se ha indicado, con el suficiente número de componentes, un GMM puede aproximar cualquier fdp. Sin embargo, a mayor número de componentes

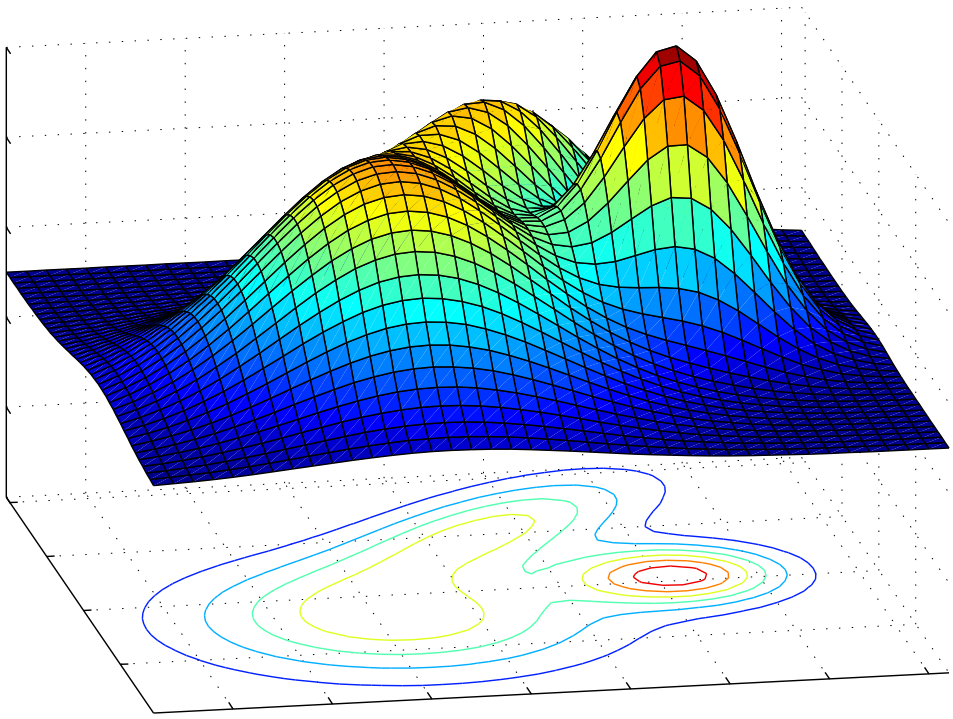


FIGURA 2.4: *Ejemplo de un GMM en un espacio bidimensional. El uso de tres componentes gaussianas permite aproximar la función de densidad de probabilidad representada por las curvas de nivel.*

gaussianas se necesitan más datos de entrenamiento para poder estimar sus parámetros. Por tanto, el número de componentes seleccionado implica un compromiso entre la precisión del modelo y su capacidad de generalización:

- Un número de componentes excesivamente bajo implica que el modelo no podrá aproximar la distribución con suficiente precisión. Estará subentrenado, por lo que aumentará la probabilidad de error del sistema.
- Si la cantidad de datos de entrenamiento no es suficiente para entrenar todas las componentes, los modelos estarán sobreentrenados, aprendiendo excesivo detalle de la base de datos de entrenamiento, y no generalizarán en muestras desconocidas.

Lo ideal por tanto, es disponer de muchas muestras de entrenamiento, de forma que se pueda utilizar un elevado número de componentes sin miedo a sobreentrenar. Debido a esto, los GMM suelen utilizarse sobre todo con parametrizaciones

segmentales, donde se extrae un vector de características por cada trama de 10-20 ms. Esto proporciona un elevado número de muestras por cada señal de entrenamiento, al contrario que las parametrizaciones supra-segmentales, que muchas veces calculan un único vector de parámetros por cada grabación. Podemos encontrar ejemplos de estos sistemas GMM aplicados a identificación de emociones con parámetros segmentales en los trabajos de Truong y van Leeuwen (2007), Vlasenko *et al.* (2007) y Kim *et al.* (2007) (ver el resumen de la Tabla 2.2).

Ververidis y Kotropoulos (2005) por el contrario utilizan GMM con estadísticos prosódicos a largo plazo. El resultado obtenido (56 % de acierto identificando 5 emociones) es muy inferior al obtenido en otros trabajos que modelan la información prosódica supra-segmental mediante otros métodos, posiblemente debido a la falta de muestras de entrenamiento. De hecho, obtienen la máxima precisión con sólo dos componentes gaussianas, incrementar más el número de componentes reduce la precisión por sobreentrenamiento. Con un número tan bajo de componentes, el modelo no es capaz de modelar adecuadamente la distribución de las emociones. También López-Cozar *et al.* (2008) utilizan GMM con estadísticos prosódicos a largo plazo, aunque en este caso combinan los resultados con los obtenidos con GMM de parámetros espectrales a nivel de trama, alcanzando un 94 % de precisión sobre dos emociones. Por desgracia no se indican los resultados parciales, por lo que no se sabe cuál es la aportación de los parámetros a largo plazo.

2.3.3. Modelos ocultos de Markov (HMM)

Los modelos ocultos de Markov (HMM, *Hidden Markov Models*) (Rabiner, 1989) pueden definirse como máquinas de estados doblemente estocásticas, en la que tanto las transiciones entre estados como los valores de emisión en cada estado se rigen por variables aleatorias. Como puede verse en la Figura 2.5, la transición entre dos estados i y j tiene una probabilidad α_{ij}, mientras que el vector de salida x emitido por el estado S_i sigue una distribución aleatoria $P_i(x)$. También el estado inicial es aleatorio, siendo π_i la probabilidad de que el estado S_i sea el inicial. Todos estos valores tienen las siguientes restricciones:

$$\alpha_{ij} \geq 0 \qquad \sum_{j=1}^{M} \alpha_{ij} = 1 \tag{2.7}$$

$$\pi_i \geq 0 \qquad \sum_{i=1}^{M} \pi_i = 1 \tag{2.8}$$

siendo M el número de estados. Se denominan modelos *ocultos* de Markov, porque mientras que la secuencia de vectores de salida $x(t)$ es conocida, la secuencia de

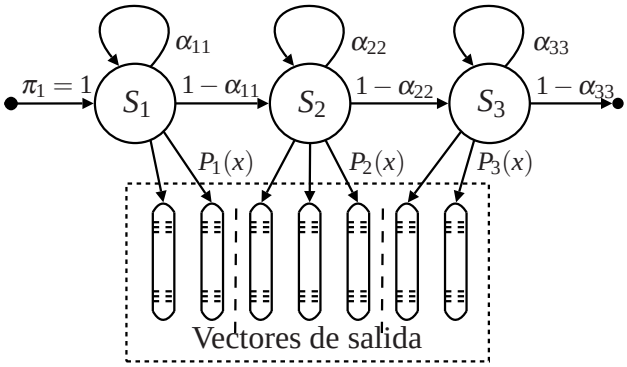


FIGURA 2.5: *Representación gráfica de un HMM de tres estados de izquierda a derecha. En cada etapa se emite un vector de salida x en función de la probabilidad de emisión del estado actual: $P_i(x)$. La secuencia de vectores de salida tiene una longitud aleatoria, pues depende del número de etapas transcurrido hasta que se sale del último estado, lo que a su vez depende de las variables aleatorias α_{ij}.*

estados por los que ha pasado el modelo no lo es. En general se permite cualquier transición entre estados. Un modelo en el que todos los estados están conectados con todos los demás se denomina *ergódico*, mientras que en los llamados *de izquierda a derecha*, como el representado en la Figura 2.5, sólo se permiten las transiciones en un sentido. En este caso no se pueden revisitar estados anteriores:

$$\alpha_{ij} = 0 \qquad \forall j < i \tag{2.9}$$

Dada una serie de secuencias de entrenamiento $x_n(t)$, el algoritmo Baum-Welch (Rabiner, 1989) permite estimar los parámetros del modelo: π_i, α_{ij} y $P_i(x)$. Puesto que en general no se tiene conocimiento previo acerca de la distribución de los vectores, $P_i(x)$ suele modelarse mediante mezclas de gaussianas, debido a su capacidad para aproximar cualquier distribución. Una vez entrenado un modelo para cada clase c, se puede clasificar una secuencia desconocida $x(t)$ calculando la probabilidad de que haya sido generada por cada uno de los modelos y seleccionando la clase correspondiente al modelo más probable:

$$C(x) = \arg\max_c P(c|x) = \arg\max_c \frac{P(x|c)P(c)}{P(x)} = \arg\max_c P(x|c)P(c) \tag{2.10}$$

donde $P(x|c)$ puede calcularse mediante el algoritmo adelante-atrás (*forward-backward*) (Rabiner, 1989).

Al contrario de los GMM, que estiman una única fdp para modelar la distribución conjunta de todos los vectores de una secuencia, los HMM manejan

las secuencias no estacionarias de forma natural. Gracias a las transiciones entre estados, permiten estimar una fdp diferente para cada sección de la secuencia, modelando así la dinámica de los parámetros. Puesto que la voz es un proceso secuencial en el tiempo, los sistemas de identificación de emociones suelen utilizar modelos de izquierda a derecha. Sin embargo también es posible utilizar modelos ergódicos, tal y como hacen Nwe *et al.* (2003).

Debido a su propia naturaleza, los HMM necesitan una secuencia de vectores para realizar el entrenamiento y la clasificación. Por ello sólo se usan con parametrizaciones a nivel segmental, habitualmente de naturaleza espectral. Sin embargo, también es posible usar HMM con las primitivas de la prosodia (valores instantáneos de F_0 y energía). Mediante esta técnica Nogueiras *et al.* (2001) han logrado un 83 % de acierto en un sistema que discrimina entre 7 emociones. En vista de este resultado se deduce que el método es efectivo, y que merece más atención por parte de los desarrolladores. Posiblemente la eficacia de esta técnica radique en que los HMM son capaces de caracterizar la forma de las curvas de entonación y energía gracias a su estructura de estados. Se trata por tanto de otra manera de modelar la información prosódica, contenida en la forma de estas curvas.

2.3.4. Vecinos más próximos (kNN)

Los modelos de *k* vecinos más próximos (kNN, *k-Nearest Neighbors*) son modelos no paramétricos que clasifican un vector x en función de las muestras de entrenamiento que queden alrededor de la misma (Duda *et al.*, 2001). Se buscan las k muestras más cercanas a x y se le asigna la clase más representada, tal y como se refleja en la Figura 2.6.

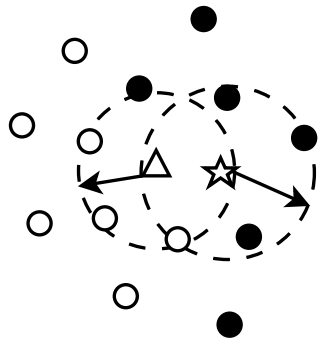


FIGURA 2.6: *Representación gráfica de un clasificador kNN con $k = 4$. La estrella se clasifica como punto negro, ya que de sus cuatro vecinos más próximos, tres son negros y uno blanco. Sin embargo, el triángulo se clasifica como punto blanco.*

El valor de *k* define la capacidad de generalización del sistema. Con un valor de *k* excesivamente pequeño, el sistema es muy sensible a las muestras cercanas, con lo que muestras ruidosas o fuera de rango pueden provocar errores cerca de la frontera. Valores mayores de *k* hacen que se busquen vecinos lejos de la muestra a clasificar, lo que conlleva una pérdida de precisión, a no ser que el número de muestras de entrenamiento sea suficientemente grande. Un número elevado de muestras de entrenamiento hace que dichas muestras estén más cerca unas de otras, con lo que la búsqueda de vecinos sigue realizándose relativamente cerca de la muestra a clasificar.

Debido a su propia naturaleza, los modelos kNN pueden utilizarse con éxito con relativamente pocos vectores de entrenamiento, dado un valor de *k* suficientemente bajo. Por ello son utilizados sobre todo con parametrizaciones suprasegmentales, y en concreto, con parámetros prosódicos. Ejemplos de clasificadores kNN utilizados con parámetros prosódicos pueden encontrarse en los trabajos de Seppänen *et al.* (2003), Kim *et al.* (2007), Lee *et al.* (2001) y Grimm *et al.* (2007).

2.3.5. Redes neuronales artificiales (ANN)

Las redes neuronales artificiales (ANN, *Artificial Neural Networks*) tratan de simular el mecanismo de funcionamiento de las neuronas cerebrales (Duda *et al.*, 2001). Como se aprecia en la Figura 2.7, cada nodo o neurona tiene una serie de entradas y una salida, de forma que el valor de esta salida es el resultado de aplicar una cierta función $f(\mathbf{a})$ al vector de entradas. La red de neuronas se compone de diferentes capas de neuronas, de forma que las salidas de una capa están conectadas a las entradas de la siguiente. También puede haber conexiones de retroalimentación, con algunas salidas conectadas a las entradas de otra capa anterior. Se ha demostrado que una red neuronal con la arquitectura adecuada es capaz de aproximar en su salida cualquier función de las entradas (Hornik *et al.*, 1989).

Cuando se utilizan para clasificación, la primera capa tiene tantas neuronas como parámetros utilizados, mientras que la última tiene tantas neuronas como clases consideradas. El algoritmo de entrenamiento estima los parámetros de la función de activación de cada neurona, procurando que cada nodo de salida se active sólo cuando a la entrada haya un vector de parámetros correspondiente a su clase. Para clasificar una muestra desconocida, se comprueba cuál de estos nodos de la última capa tiene el mayor valor de salida, asignándole la clase correspondiente.

A mayor número de neuronas mayor capacidad tiene el sistema para modelar salidas complejas, lo que permite ganar precisión a la hora de calcular las fronteras entre las clases. Sin embargo, también se necesitan más datos de entrenamiento

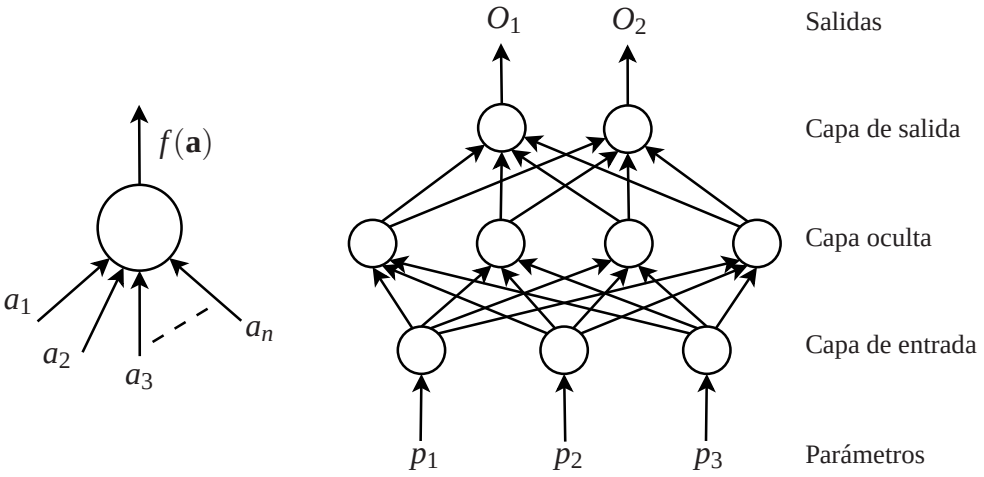


FIGURA 2.7: *Izquierda: Representación gráfica de una neurona con n entradas a_i. El valor de salida $f(\mathbf{a})$ es una función de todas las entradas. Derecha: Esquema de una red neuronal de tres capas. Toma tres parámetros como entrada y genera dos salidas O_1 y O_2. Esta ANN puede utilizarse como un clasificador de dos clases utilizando tres parámetros.*

para estimar todos los parámetros necesarios, con lo que existe peligro de sobreentrenamiento. Por el contrario, un número muy reducido de neuronas supone que el sistema no será capaz de aprender fronteras complejas con suficiente precisión.

La mayoría de los trabajos de identificación de emociones que hacen uso de ANN las aplican para modelar parámetros prosódicos supra-segmentales. Es el caso de Jiang y Cai (2004), Hozjan y Kacic (2003), Yacoub *et al.* (2003), Tato *et al.* (2002) y Steidl *et al.* (2005). Aunque menos habitual, también se pueden encontrar trabajos que utilizan ANN con parámetros segmentales a corto plazo, como el de Nicholson *et al.* (2000), que utiliza una combinación de coeficientes de predicción lineal (LPC, *Linear Prediction Coefficients*) y valores instantáneos de entonación y potencia.

Petrushin (2000) ofrece una comparación de los resultados obtenidos con ANN y kNN sobre el mismo experimento. Mientras que con ANN se obtiene un 65 % de precisión, los kNN sólo llegan al 55 %. También el trabajo de Yacoub *et al.* (2003) presenta una comparativa, esta vez entre ANN y SVM. En el sistema presentado, los ANN obtienen un 94 % de acierto frente al 91 % de las SVM. Sin embargo, al reducir el número de datos de entrenamiento, son las SVM las que obtienen mejor precisión: un 91 % frente al 87 % de los ANN. Es interesante ver cómo el resultado de las SVM es idéntico mientras que las ANN han incrementado el error al doble.

2.3.6. Máquinas de vectores soporte (SVM)

Las máquinas de vectores soporte (SVM, *Support Vector Machines*) (Vapnik, 1995) han tenido un gran auge en los últimos años debido a su gran capacidad de generalización y la posibilidad de conseguir buenos resultados de clasificación incluso cuando se dispone de muy pocos datos de entrenamiento.

Aunque los conceptos básicos de una SVM son bastante simples, el desarrollo matemático de los mismos hasta alcanzar la solución del sistema es muy largo. En esta sección sólo se describen las propiedades principales de las SVM, sin entrar en detalles matemáticos. En caso de tener interés por estos detalles, pueden encontrarse en la extensa bibliografía dedicada al tema. Pueden destacarse los artículos de Osuna *et al.* (1997), Burges (1998), Chen *et al.* (2005) y Sanchez A. (2003) por su claridad y profundidad.

Las SVM se basan en un concepto simple pero eficaz: de todas las posibles fronteras que separan dos clases, se busca aquella que maximiza la distancia entre los datos de entrenamiento y la frontera. Esta distancia se denomina *margen*. Puesto que la mayor probabilidad de error se da precisamente en la región cercana a la frontera, al maximizar el margen se logra minimizar el error. La Figura 2.8 representa esta idea con un ejemplo gráfico. En vista de la distribución de las muestras de entrenamiento, el aspa parece pertenecer a la clase de círculos blancos. Sin embargo una solución con poco margen puede clasificar este punto en el lado contrario de la frontera.

En su forma más básica, las SVM sólo pueden aplicarse a problemas de dos clases. Sean $C = \{+1, -1\}$ las etiquetas asociadas a esas clases. Entonces la frontera se calcula como un hiperplano $\mathbf{w} \cdot \mathbf{x} + b = 0$, y la clasificación de un nuevo vector de parámetros $\mathbf{x}$ se realiza mediante:

$$\hat{c} = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \tag{2.11}$$

Aplicando la condición de maximizar el margen puede llegarse a la expresión:

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x_i} \tag{2.12}$$

donde N es el número de muestras de entrenamiento, $\mathbf{x_i}$ son las muestras, $y_i \in C$ son sus etiquetas de clase correspondientes y λ_i es un valor asociado a cada muestra de entrenamiento que hay que calcular, generalmente mediante un algoritmo de programación cuadrática. Sólo las muestras de entrenamiento para las que $\lambda_i \neq 0$ son necesarias para el cálculo de la frontera, y estas muestras se denominan *vectores soporte*. En el caso linealmente separable (Figura 2.8(a)) estas muestras resultan ser precisamente las que están justo encima del margen. En el caso no separable, en el que se permite que muestras de entrenamiento caigan dentro del

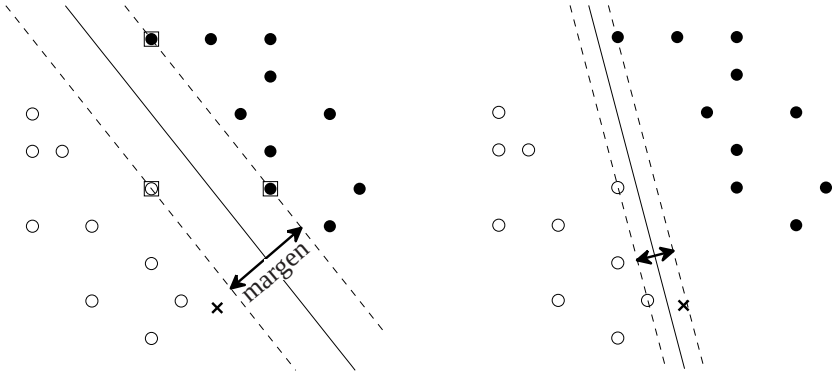


FIGURA 2.8: *Ejemplo de una SVM aplicada a un problema de dos clases linealmente separables. (a) La SVM busca la frontera que maximiza el margen, con lo que se reduce la probabilidad de error sobre muestras no vistas. Los vectores soporte figuran encuadrados. (b) Aunque existen infinitas fronteras que consiguen separar las muestras de entrenamiento, proporcionan menor generalización y más error sobre muestras no vistas, debido a la reducción del margen. Probablemente el aspa pertenece a la clase de círculos blancos, sin embargo el sistema (b) lo clasifica en la clase de círculos negros.*

margen o estén incluso mal clasificadas, son vectores soporte todas aquellas muestras que caigan encima del margen, dentro del mismo, o que estén al otro lado de la frontera. Es decir, la forma de la frontera sólo se ve afectada por las muestras más cercanas a la clase vecina. Eliminar cualquier muestra que no sea un vector soporte no varía la solución. Dicho de otra manera, las SVM calculan la frontera con las muestras más conflictivas, que son las que pueden dar lugar a errores de clasificación, y se despreocupan de las muestras alejadas.

Con el uso de kernels (Müller *et al.*, 2001; Schölkopf y Smola, 2001) las *SVM* han evolucionado para permitir fronteras no lineales. La idea es hallar una transformación no lineal Φ:

$$\begin{array}{rcl} \Phi : \mathbb{R}^N & \rightarrow & \mathscr{F} \\ \mathbf{x} & \mapsto & \Phi(\mathbf{x}) \end{array} \tag{2.13}$$

que transforme los vectores $\mathbf{x} \in \mathbb{R}^N$ a un espacio de parámetros $\mathscr{F}$ de mayor dimensión. Si en este espacio de parámetros las clases sí son linealmente separables, se pueda aplicar la SVM. Gracias al uso de kernels no es necesario definir esta transformación Φ explícitamente, basta con hallar una función (kernel) $k(\mathbf{x}, \mathbf{y})$ tal que:

$$k(\mathbf{x},\mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \tag{2.14}$$

Es decir, el resultado de aplicar el kernel a dos vectores es igual al producto escalar de esos dos vectores en el espacio transformado. Puesto que es posible describir una SVM sólo en función de productos escalares, sustituyendo estos productos por un kernel se obtiene el mismo resultado sin tener que aplicar la transformación a los vectores de forma explícita. Por ejemplo, la función de decisión (2.11) puede escribirse como:

$$\hat{c} = \text{sign}(k(\mathbf{w},\mathbf{x}) + b) \tag{2.15}$$

Los kernels más utilizados son:

Kernel RBF	$k(\mathbf{x},\mathbf{y}) = \exp\left(\frac{-\|\|\mathbf{x}-\mathbf{y}\|\|^2}{2\sigma^2}\right)$
Kernel polinómico	$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+b)^d$
Kernel sigmoide	$k(\mathbf{x},\mathbf{y}) = \tanh(\mathbf{x}\cdot\mathbf{y}+\theta)$

Los parámetros σ, b, d y θ de estos kernels están relacionados con la dimensión del espacio transformado. Es necesario estimarlos con cuidado ya que tienen una influencia directa en la separabilidad del problema en ese espacio, y por tanto, en la capacidad de generalización del sistema. La optimización de estos parámetros es un tema recurrente en la literatura (Ayat *et al.*, 2005; Chapelle *et al.*, 2002; Keerthi y Lin, 2003; Schittkowski, 2005; Wang *et al.*, 2003; Wu y Wang, 2008), aunque generalmente sus valores se fijan mediante pruebas de validación cruzada.

La Figura 2.9 muestra un ejemplo práctico utilizando un kernel RBF. Las clases representadas no son linealmente separables en el espacio de parámetros considerado. Sin embargo, utilizando un kernel con un valor de σ apropiado se puede hallar una frontera no lineal más adecuada para el problema (Figura 2.9(a)). Nótese que en este caso el problema tampoco es separable en el espacio transformado, pues quedan muestras de entrenamiento mal clasificadas.

Variando el valor de σ se modifica la frontera calculada. Un valor de σ muy elevado suaviza la frontera, dando como resultado una solución parecida a la que proporciona una SVM lineal (Figura 2.9(b)). Un valor pequeño permite que la frontera se adapte mejor a los datos de entrenamiento, con el consiguiente peligro de sobreentrenamiento y mala generalización (Figura 2.9(c)).

Este comportamiento se explica al comprobar cómo afecta el valor de σ en el cálculo de la función de decisión (2.15). Cada muestra de entrenamiento influye en el espacio vectorial a su alrededor tratando de asignarlo a su clase. El tamaño del entorno sobre el que influye una muestra viene determinado por σ: a mayor

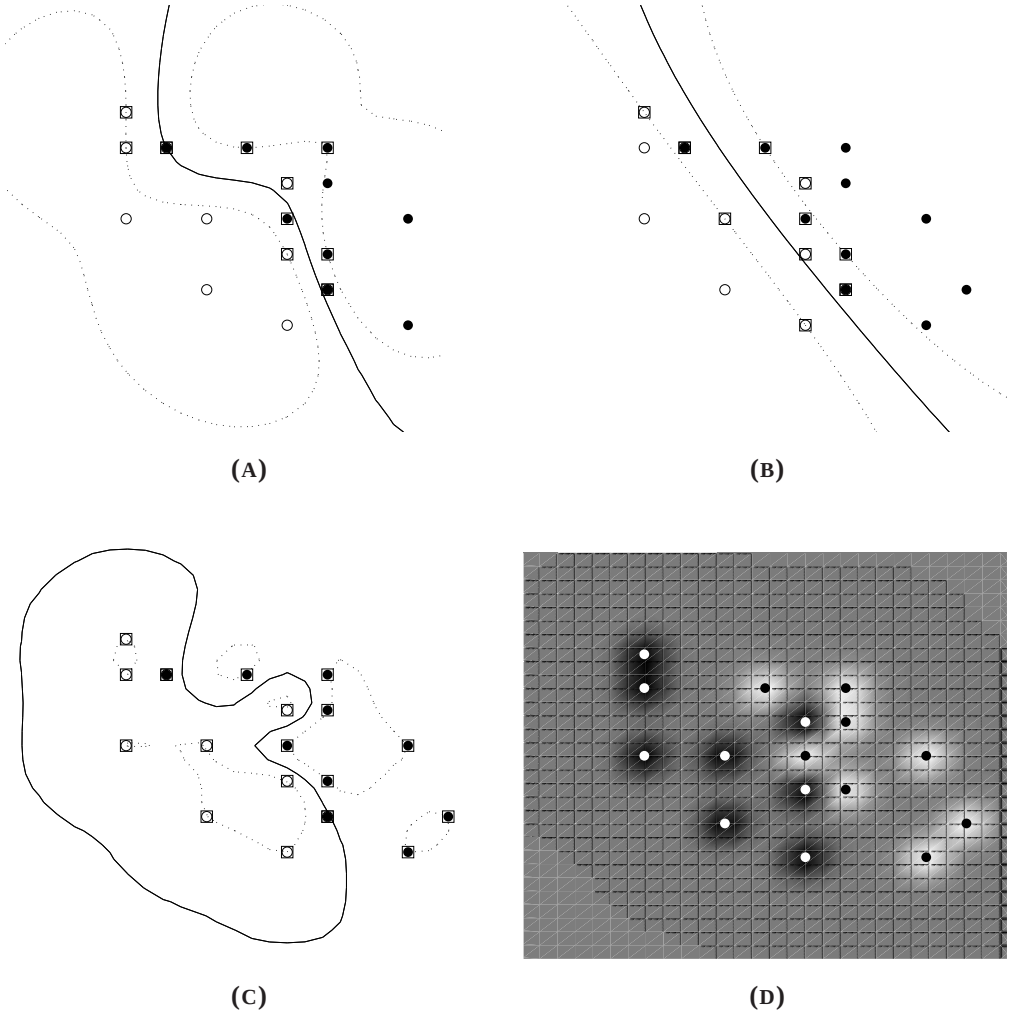


FIGURA 2.9: *Ejemplo de una SVM con kernel RBF. (a) La utilización del kernel permite fronteras no lineales. También se observa que en un caso no separable todos los vectores dentro del margen son vectores soporte. (b) Con un valor de σ grande el sistema converge a una frontera lineal. (c) Con un valor de σ excesivamente pequeño la frontera se sobreentrena: casi todas las muestras de entrenamiento se clasifican correctamente (error empírico reducido) pero el sistema generaliza mal. Todas las muestras de entrenamiento se convierten en vectores soporte, lo que ayuda a detectar el problema. (d) Representación del valor de la función de decisión con σ pequeño. Los colores claros indican un valor positivo mientras que los oscuros representan valores negativos. Puede verse que las zonas alejadas de cualquier vector de entrenamiento tienen un color gris intermedio.*

valor de σ, mayor es el entorno afectado. Con valores muy pequeños sólo el entorno inmediato de cada muestra está claramente posicionado en una u otra clase, mientras que el resto queda a medio camino entre las dos. La consecuencia es que **todos** los vectores acaban siendo vectores soporte, y la mayor parte del espacio vectorial queda dentro del margen. De ahí la poca generalización del sistema: cualquier muestra nueva un poco alejada de las de entrenamiento se sitúa en una zona dudosa. Este efecto puede verse en la Figura 2.9(d), donde se representa el valor de la función $k(\mathbf{w},\mathbf{x})+b$ para una σ muy pequeña. Como puede apreciarse, la mayor parte del espacio tiene un valor intermedio. Por el contrario, con un valor de σ grande el entorno de influencia de cada muestra es mayor. Esto provoca que cada punto del espacio se vea afectado por varias muestras de entrenamiento, promediando el efecto de las muestras positivas y negativas y suavizando la frontera.

Por definición las SVM sólo permiten separar entre dos clases. Sin embargo, se han desarrollado métodos para permitir aplicarlas sobre problemas multiclase. Estos métodos consisten en dividir el problema en varios sub-problemas de dos clases y combinar los resultados. Mayoritariamente se utilizan los métodos uno-contra-todos y uno-contra-uno, aunque se han propuesto otras divisiones alternativas.

El método uno-contra-todos divide el problema de N clases en N problemas de dos clases, de forma que cada sub-problema separa una de las clases de todas las demás. La clasificación se realiza asignando la clase que obtenga un mayor valor en la función de decisión. El método uno-contra-uno por el contrario separa todas las posibles parejas de clases entre sí, entrenando por tanto $\frac{N(N-1)}{2}$ sub-sistemas. En este caso, la decisión final suele hacerse por voto por mayoría. Otras alternativas propuestas tratan de optimizar el número y tipo de sub-problemas. Por ejemplo, estructurando las clases en una jerarquía en forma de árbol binario se puede entrenar una SVM en cada nodo para separar las dos ramas descendientes, lo que requiere $\log_2(N)$ sub-sistemas (Madzarov *et al.*, 2009). Chen *et al.* (2009) y Lee *et al.* (2009) presentan otros sistema similares, también basados en árboles binarios. Hsu y Lin (2002) realizan una revisión de varias alternativas para la clasificación multiclase y comparan su rendimiento y precisión.

Como ya se ha comentado en la sección 2.3.5, el trabajo de Yacoub *et al.* (2003) pone de manifiesto la capacidad de las SVM a la hora de trabajar con pocos datos de entrenamiento y su gran capacidad de generalización. Por tanto, son muy adecuadas para modelar parámetros supra-segmentales. Es el caso de Morrison *et al.* (2007) y Vlasenko *et al.* (2007). Muchos autores aprovechan la flexibilidad de las SVM para combinar varios tipos de parámetros, tal y como hacen Müller *et al.* (2004) (combinando parámetros prosódicos y lingüísticos), Kwon *et al.* (2003) (prosodia y estadísticos de espectro) y Schuller *et al.* (2005) (prosodia, estadísticos de espectro y parámetros lingüísticos). Sin embargo, tam-

bién hay intentos de utilizar las SVM con parámetros a nivel de trama, como el realizado por Shami y Verhelst (2007), que utilizan parámetros espectrales y valores de F_0 y energía.

2.3.7. Conclusiones del análisis de los clasificadores

La elección del clasificador viene condicionada en gran medida por el tipo de parámetros utilizados y el número de clases a discriminar. Debido a su naturaleza y a cómo trata los datos de entrada, cada clasificador funciona mejor en un tipo determinado de parametrización. Los sistemas que modelan la distribución de los parámetros o *modelos generativos* (GMM, HMM) necesitan muchos datos de entrenamiento, por lo que dan mejores resultados con los parámetros a nivel de trama. Por el contrario, los *modelos discriminativos* (kNN, SVM) parecen funcionar mejor con parámetros supra-segmentales. Por tanto, no es de extrañar que muchos autores, enfrentados con el problema de utilizar tanto parámetros segmentales como supra-segmentales, decidan utilizar diferentes modelos para cada tipo de parametrización, y combinar posteriormente los resultados mediante un sistema de fusión tardía (Kim *et al.*, 2007; López-Cozar *et al.*, 2008; Müller *et al.*, 2004; Vlasenko *et al.*, 2007).

2.4. Conclusiones

No es sencillo realizar un estudio comparativo de los sistemas de identificación de emociones publicados en la literatura. Entre otras razones, porque son pocos los trabajos publicados que utilizan una misma base de datos común, con lo que los resultados no son comparables debido a las diferencias en el contenido de estas bases de datos (Tabla 2.1): el número y tipo de emociones, origen de las mismas (naturales, evocadas, estimuladas o actuadas), número de locutores, contenido de las grabaciones (textos leídos o voz espontánea, frases cortas, largas o palabras aisladas), etc.

Es cierto que algunas bases de datos puestas a disposición del público han sido utilizadas más frecuentemente, facilitando cierta comparación, como por ejemplo *Berlin* (Burkhardt *et al.*, 2005) o *SUSAS* (Hansen y Bou-Ghazale, 1997). Por desgracia el número de sistemas comparables no es suficiente como para poder extraer conclusiones definitivas. Existen, además de la base de datos, otras diferencias que también dificultan la comparación, como puede ser la dependencia o independencia del locutor de las pruebas realizadas.

Por tanto pueden identificarse tres variables principales a la hora de describir un trabajo de identificación de emociones:

- El **número de emociones**. A mayor número de emociones mayor confusión se introduce en el sistema, aumentando la probabilidad de error.
- La **naturaleza de las emociones**. Los experimentos realizados sobre una base de datos de emociones naturales reflejan peores resultados que los realizados sobre una base de datos actuada, posiblemente debido a la sobreactuación en estas últimas.
- La **dependencia con el locutor**. Las pruebas realizadas mediante una arquitectura dependiente de locutor, es decir, cuando el locutor de las señales de prueba forma parte de la base de datos de entrenamiento, proporcionan mejores resultados, ya que el sistema ha tenido la oportunidad de aprender las características de la voz de ese locutor. Sin embargo, estos resultados no son extrapolables a una implementación comercial, donde los locutores que utilizan el sistema son desconocidos.

La grabación y publicación de *AIBO* (Batliner *et al.*, 2006), una base de datos con un número suficiente de locutores y habla espontánea reflejando emociones naturales, puede suponer un cambio en este sentido. Sobre todo después de la celebración del *Emotion Challenge* (Schuller *et al.*, 2009) durante la conferencia Interspeech'09, una competición de identificación de emociones en la que diferentes grupos de investigación competían con sus propios sistemas tomando como base de datos *AIBO*. La organización del evento proporciona además los conjuntos de entrenamiento y evaluación asegurando una arquitectura independiente de locutor. Este evento puede establecer esta base de datos como referencia a la hora de comparar y evaluar los sistemas, al hacerse públicos los resultados obtenidos por los grupos de investigación participantes (Barra-Chicote *et al.*, 2009; Bozkurt *et al.*, 2009; Lee *et al.*, 2009; Luengo *et al.*, 2009a; Polzehl *et al.*, 2009; Vogt y André, 2009).

En el caso de las parametrizaciones, existen además factores añadidos que complican la comparación de los resultados. Incluso aunque se utilice una misma base de datos y arquitectura de pruebas, habitualmente se emplean diferentes clasificadores y número de parámetros para cada parametrización. Bajo estas condiciones no es posible determinar si una mejora de los resultados es debida al uso de una parametrización diferente o a haber utilizado un mayor número de parámetros o un clasificador específico.

En esta revisión del estado del arte se ha intentado por lo menos mostrar algunos ejemplos representativos para proporcionar una idea general de las precisiones obtenidas en cada caso y las alternativas propuestas para cada una de las etapas de diseño del sistema (base de datos, parametrización y clasificación). La Tabla 2.2 proporciona una visión de los sistemas descritos, su arquitectura y los resultados obtenidos.

TABLA 2.2: *Resumen de algunos trabajos sobre identificación de emociones.*

Tipo de emociones: A *(actuadas)*, N *(naturales)*.
Arquitectura de las pruebas: IL *(independiente de locutor)*, DL *(dependiente de locutor)*, LU *(locutor único)*.
Tipo de parametrización: LP *(prosodia a largo plazo)*, SP *(prosodia a corto plazo)*, LS *(espectro a largo plazo)*, SS *(espectro a corto plazo)*, VQ *(calidad de voz)*, L *(lingüísticos)*.

Referencia	#Emo	A/N	IL/DL/LU	Parámetros	#Par	Clasificador	Resultado	Observaciones
McGilloway *et al.* (2000)	5	A	DL	LP	375	LDA SVM	55% 52%	
Nicholson *et al.* (2000)	8	A	IL	SP+SS	15	ANN	55%	
Petrushin (2000)	5	A	DL	LP	43	kNN ANN	55% 65%	
Lee *et al.* (2001)	2	N	DL	LP	10	kNN/LDA	75%(hombres) 80%(mujeres)	Realiza el experimento por separado para hombres y mujeres.
Nogueiras *et al.* (2001)	7	A	DL	SP	4	HMM	83%	
Tato *et al.* (2002)	5	N	IL	LP+VQ	53	ANN	60%	
Hozjan y Kacic (2003)	7	A	LU	LP	144	ANN	60-90%	Base de datos *Interface*. El resultado depende del locutor utilizado.
Kwon *et al.* (2003)	5 4 2	A N N	DL	LP+LS	59	SVM	42% 67% 91%	Bases de datos *AIBO* (5 emo.) y *SUSAS* (4 y 2 emo.).
Nwe *et al.* (2003)	6	A	LU	SS	12	HMM	75-80%	El resultado depende del locutor considerado.
Pierre-Yves (2003)	4	A	DL	LP+LS	200	Varios	80-95%	El resultado depende en gran medida del clasificador.
Seppänen *et al.* (2003)	4	A	IL DL	LP	43	kNN	60% 75%	
Yacoub *et al.* (2003)	2	A	IL	LP	37	SVM ANN	91% 94%	

Continúa en la siguiente página

TABLA 2.2: *Continuación.*

Referencia	#Emo	A/N	IL/DL/LU	Parámetros	#Par	Clasificador	Resultado	Observaciones
Jiang y Cai (2004)	6	A	LU	LP	8	ANN	94%	
Müller *et al.* (2004)	7	A	-	LP+L	200	SVM	92%	74% sólo con parámetros acústicos. 60% sólo con parámetros lingüísticos.
Luengo *et al.* (2005)	7	A	LU	LP SS	86 36	SVM GMM	92% 98%	
Schuller *et al.* (2005)	7	A	IL	LP+LS+L	276	SVM	88%	
Steidl *et al.* (2005)	4	N	IL	LP+L	125	ANN	60%	
Ververidis y Kotropoulos (2005)	5	A	DL	LP	87	GMM	56%	
Vogt y André (2005)	7	A	DL	LP+LS	1280	Redes Bayesianas	77%	
Vogt y André (2006)	7	A	IL	LP+LS	1280	Redes Bayesianas	81%	
Batliner *et al.* (2006)	4	N	IL	Varios	Varios	Varios	50-60%	Recoge los resultados de varias sedes sobre una misma base de datos, utilizando cada sede su propio sistema.
Casale *et al.* (2007)	4 2	A	DL	SS	462	HMM	83% 95%	Consigue un 45% con sólo MFCC con 4 emociones, y un 77% con 2 emociones.
Chichosz y Slot (2007)	7	A	IL DL	LP+LS	102	Árboles	72% 74%	
Grimm *et al.* (2007)	4	A	IL DL	LP+LS	46	kNN	67% 84%	

Continúa en la siguiente página

TABLA 2.2: *Continuación.*

Referencia	#Emo	A/N	IL/DL/LU	Parámetros	#Par	Clasificador	Resultado	Observaciones
Kim *et al.* (2007)	2	A	IL	LP+SS	12 (LP)	kNN/GMM	83-95%	83% con señales de 1 segundo. 95% con señales de 5 segundos. Utiliza 12 parámetros prosódicos y MFCC para el espectro, aunque no indica cuántos ni si calcula diferencias.
Lugger y Yang (2007)	6	A	IL	LP+VQ	208	Redes Bayesianas	75%	67% sólo con parámetros prosódicos. 61% sólo con calidad de voz.
Morrison *et al.* (2007)	6 2	A N	DL	LP+LS	38	SVM	72% 79%	Utiliza una base de datos actuada de 6 emociones y otra natural de 2 emociones.
Shami y Verhelst (2007)	7 5 5 3	A A A N	DL	LP+LS	560	SVM	76% 64% 84% 66%	Utiliza 4 bases de datos: *Berlin* (7 emociones actuadas), *Danish* (5 emociones actuadas), *Kismet* (5 emociones actuadas) y *BabyEars* (3 emociones naturales)
Truong y van Leeuwen (2007)	7	A	IL	SS	26	GMM	75%	
Vlasenko *et al.* (2007)	7 2	A N	DL	LP+SS	1406	SVM/GMM	90% 84%	Utiliza la base de datos *Berlin* (7 emociones actuadas) y *SUSAS* (2 emociones naturales).
López-Cozar *et al.* (2008)	2	N	DL	LP+SS+L	-	GMM	95%	
Barra-Chicote *et al.* (2009)	5 2	N	IL	SP+SS	6+39	Redes Bayesianas	38% 67%	
Bozkurt *et al.* (2009)	5 2	N	IL	SP+SS	3+87	GMM	41% 68%	Sólo con prosodia: 34% (5 emo.) y 63% (2 emo.). Sólo con espectro: 41% (5 emo.) y 68% (2 emo.).
Lee *et al.* (2009)	5	N	IL	LP+LS	384	LR SVM	42% 41%	
Luengo *et al.* (2009a)	5 2	N	IL	LP+SS	56+54	SVM/GMM	41% 67%	Sólo con prosodia: 35% (5 emo.) y 60% (2 emo.). Sólo con espectro: 39% (5 emo.) y 61% (2 emo.).

Continúa en la siguiente página

TABLA 2.2: *Continuación.*

Referencia	#Emo	A/N	IL/DL/LU	Parámetros	#Par	Clasificador	Resultado	Observaciones
Polzehl *et al.* (2009)	2	N	IL	LP+LS+VQ+L	1500	SVM	68%	Con LP+LS+VQ: 65%. Sólo con L: 68%. Con fusión tardía: 68%.
Vogt y André (2009)	5 2	N	IL	LP+LS+VQ	1451	Naïve Bayes	36% 66%	

Capítulo 3

Parámetros para la identificación de emociones en el habla

Índice

EL análisis llevado a cabo en este trabajo se centra en los parámetros de naturaleza acústica: características espectrales, prosódicas y de calidad de voz. Este tipo de parámetros son los más utilizados en los sistemas de identificación automática de emociones en el habla, ya que pueden calcularse diréctamente a partir de la señal de voz sin necesidad de conocer el contenido lingüístico de las locuciones. Además, las parametrizaciones acústicas proporcionan una mayor independencia del idioma, con lo que las conclusiones obtenidas en su análisis serán más generalizables. Este capítulo describe los parámetros considerados para este análisis y el procedimiento utilizado para su cálculo.

En la primera sección se detalla el procesado de las señales para la extracción de las curvas y etiquetas necesarias para el posterior cálculo de los parámetros. Este procesado permite calcular las marcas de actividad vocal, la señal glotal, las curvas de entonación junto con la información de sonoridad, las marcas a período de pitch y la posición estimada de las vocales. Algunos de estos procedimientos han sido desarrollados especialmente para la elaboración de esta tesis. En la sección 3.2 se detallan los parámetros considerados en la evaluación, que son extraídos a partir de la señal de voz y de los resultados del anterior procesado.

3.1. Procesado de las señales de voz

3.1.1. Estimación de la actividad vocal (VAD)

La detección de actividad vocal (VAD, *Voice Activity Detection*) trata de identificar las regiones de la señal en las que hay voz (actividad vocal) y en las que no. Esta detección permite descartar los segmentos de silencio durante el procesado de las señales. La mayoría de los algoritmos de VAD estándares se han desarrollado para aplicaciones de codificación (ETSI, 1997; ITU-T, 2007) o RAH (ETSI, 2003). En estas aplicaciones es importante que no haya segmentos de voz clasificados como silencio, ya que se perdería parte del mensaje. Sin embargo, para la detección automática de emociones es más importante minimizar el número de

tramas de silencio que se clasifican como voz. Puesto que las tramas de silencio no proporcionan información acústica acerca de la emoción, corrompen la distribución de los parámetros calculados, aumentando la confusión del sistema. Sin embargo, que algunas tramas de voz se clasifiquen como silencio no es preocupante. Es más, estas tramas probablemente tengan un nivel de energía muy bajo o estarán corruptas por ruido, por lo que proporcionarían una estimación pobre de la distribución de los parámetros. De esta forma, un incremento moderado del número de tramas de voz clasificadas como silencio puede incluso ser beneficioso.

Para la detección se ha aplicado una versión modificada del algoritmo presentado por Ramirez *et al.* (2004), el cual está basado en el cálculo de la divergencia espectral a largo plazo (LTSD, *Long-Term Spectral Divergence*) entre los segmentos que tienen actividad vocal y los que no. Las Figuras 3.1 y 3.2 pueden ser útiles para entender la descripción del algoritmo, ya que muestran el resultado de las distintas etapas del mismo. En la Figura 3.1 se ha realizado la detección en una señal limpia, con una relación señal a ruido (SNR, *Signal to Noise Ratio*) media de 20 dB, mientras que la Figura 3.2 presenta los resultados para la misma señal con una SNR media de 5 dB.

Las marcas de actividad vocal obtenidas son además post-procesadas para eliminar silencios excesivamente cortos (de menos de 100 ms), que suelen aparecer como consecuencia de errores en la detección, sobre todo en señales ruidosas. El efecto de este post-procesado puede apreciarse en la Figura 3.2, comparando la primera estimación del algoritmo (Figura 3.2c) y las marcas finales después del post-procesado (Figura 3.2a).

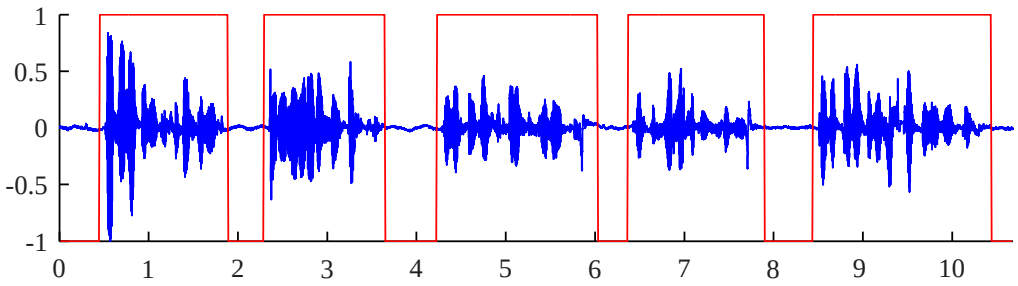
Sea $s[n]$ la señal a procesar, la cual se enventana con ventanas de tamaño fijo, generalmente solapadas, para formar L tramas $x(l)$ con $l = 1 \dots L$. Sea $X(k,l)$ la amplitud del espectro para la banda k de la trama l, calculada mediante una transformada discreta de Fourier (DFT, *Discrete Fourier Transform*), con $k = 1 \dots K$. La envolvente espectral a largo plazo (LTSE, *Long-Term Spectral Envelope*) de orden N para la trama $x(l)$ se define como:

$$LTSE(k,l) = \max_{-N \leq j \leq N} \{X(k,l+j)\} \tag{3.1}$$

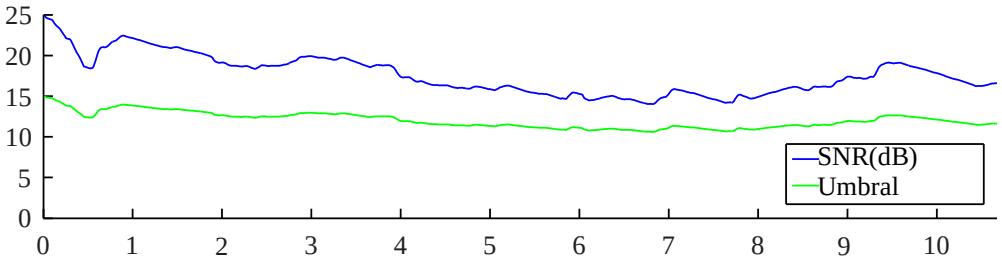
La LTSD entre una trama y el ruido se define como la desviación entre la LTSE de la trama y el espectro estimado del ruido $N(k)$:

$$LTSD(l) = 10 \cdot log_{10} \left(\frac{1}{K} \sum_{k=1}^{K} \frac{LTSE^2(k,l)}{N^2(k)} \right) \tag{3.2}$$

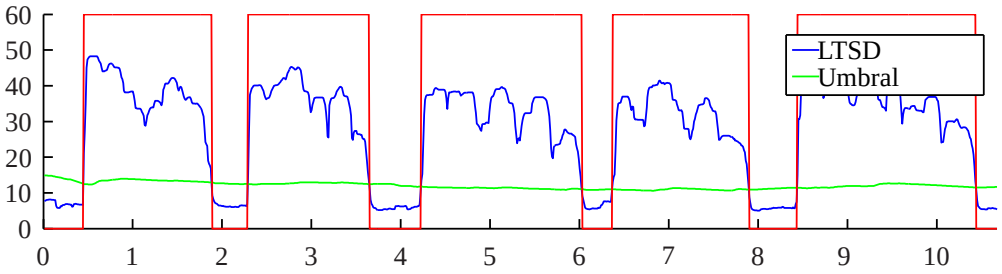
La estimación del espectro de ruido $N(k)$ y de la potencia de ruido P_N se puede realizar en una etapa de inicialización, promediando T tramas sin actividad vocal.



(A) *Forma de onda y decisión final.*



(B) *SNR y umbral.*



(C) *LTSD, umbral y primera decisión.*

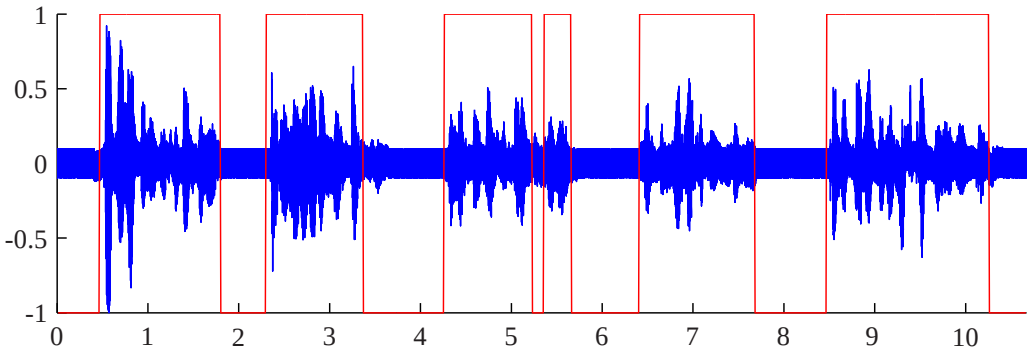
FIGURA 3.1: *Funcionamiento del VAD en una señal limpia (SNR ≈ 20 dB).*

Por ejemplo, si se sabe que hay un silencio inicial suficientemente largo, pueden usarse las T primeras tramas.

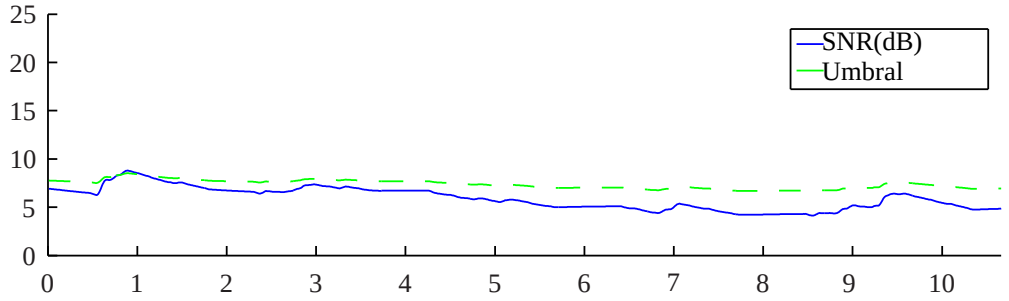
$$N(k) = \frac{1}{T}\sum_{l=1}^{T} X(k,l) \tag{3.3}$$

$$P_N = \frac{1}{KT}\sum_{l=1}^{T}\sum_{k=1}^{K} X^2(k,l) \tag{3.4}$$

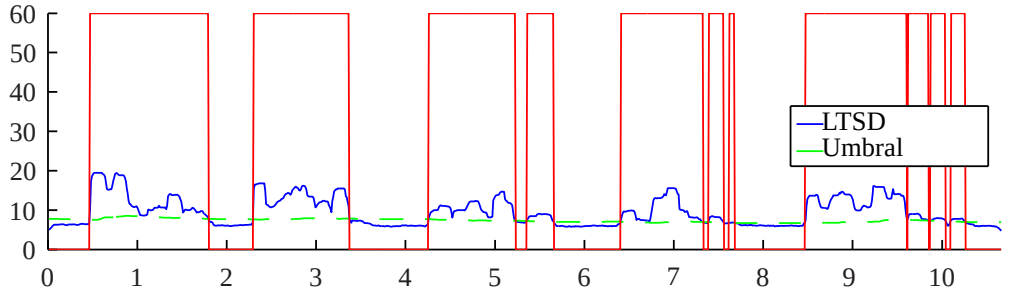
La Figura 3.1c muestra la evolución de la LTSD para la señal del ejemplo.



(A) *Forma de onda y decisión final.*



(B) *SNR y umbral.*



(C) *LTSD, umbral y primera decisión.*

FIGURA 3.2: *Funcionamiento del VAD en una señal ruidosa (SNR $\approx$ 5 dB).*

Puede apreciarse que la LTSD es mayor en zonas de actividad vocal y menor en los silencios. Si el valor de la LTSD en una cierta trama es suficientemente grande, se puede considerar que su envolvente espectral es muy diferente al espectro estimado del ruido, y etiquetarse como voz. En caso contrario, se etiqueta como silencio. Para ello se compara la LTSD con un umbral γ:

$$LTSD(l) \begin{cases} \geq \gamma & x(l) \text{ etiquetada como voz} \\ < \gamma & x(l) \text{ etiquetada como silencio} \end{cases} \tag{3.5}$$

La capacidad de un sistema VAD a la hora de detectar la actividad vocal depende de la SNR. A menor SNR, más similares son la potencia y el espectro de las regiones con y sin actividad vocal, por lo que es más fácil confundirlas. Esto quiere decir que a valores bajos de SNR el umbral γ debe ser pequeño, permitiendo que el algoritmo distinga diferencias más sutiles. El sistema original propuesto por Ramirez *et al.* (2004) fija el umbral γ en función del valor P_N estimado durante el proceso de inicialización mediante (3.4). Sin embargo, con esta aproximación se está suponiendo que la potencia de ruido permanece constante a lo largo de la señal. Además, no se tiene en cuenta la potencia de la señal a la hora de fijar el umbral. Por ello se ha modificado este algoritmo utilizando un umbral adaptativo con la SNR estimada para cada trama:

$$\gamma(l) = \begin{cases} \gamma_m & SNR(l) \leq SNR_m \\ \frac{\gamma_m - \gamma_M}{SNR_m - SNR_M}(SNR(l) - SNR_m) + \gamma_m & SNR_m < SNR(l) < SNR_M \\ \gamma_M & SNR(l) \geq SNR_M \end{cases} \tag{3.6}$$

donde $SNR(l)$ es el valor de la SNR estimada en la trama actual, SNR_m y SNR_M son la mínima y la máxima SNR consideradas y γ_m y γ_M son los umbrales predefinidos para estos niveles de SNR respectivamente. La Figura 3.3 muestra la variación del umbral γ en función de $SNR(l)$. En la Figura 3.1b se ha representado el valor de γ junto a la SNR estimada, mientras que la Figura 3.1c muestra el valor de γ junto a la LTSD, de forma que puede apreciarse la decisión tomada por el detector en cada trama.

Con el objetivo de que el detector pueda operar correctamente con señales en las que varía el nivel de ruido, se ha implementado un mecanismo adaptativo para la estimación de la SNR y la potencia de ruido P_N. Con cada trama que se clasifica como silencio, se actualiza la estimación del espectro de ruido $N(k,l)$ y su potencia $P_N(l)$ usando un factor de olvido α_N:

$$N(k,l) = \begin{cases} \alpha_N \cdot N(k,l-1) + (1-\alpha_N) \cdot X(k,l) & \text{Si silencio} \\ N(k,l-1) & \text{Si voz} \end{cases} \tag{3.7}$$

$$P_N(l) = \begin{cases} \alpha_N \cdot P_N(l-1) + (1-\alpha_N) \cdot P_X(l) & \text{Si silencio} \\ P_N(l-1) & \text{Si voz} \end{cases} \tag{3.8}$$

siendo $P_X(l)$ la potencia de la trama l. Los valores iniciales $N(k,0)$ y $P_N(k,0)$ son los obtenidos durante la inicialización mediante las ecuaciones (3.3) y (3.4). De

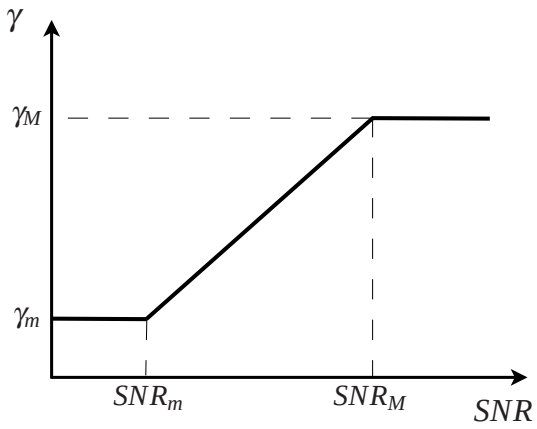


FIGURA 3.3: *Variación del umbral γ de la LTSD en función de la SNR estimada.*

forma similar, se actualiza la potencia de la señal con cada trama que se clasifica como voz:

$$P_S(l) = \begin{cases} \alpha_S \cdot P_S(l-1) + (1-\alpha_S) \cdot P_X(l) & \text{Si voz} \\ P_S(l-1) & \text{Si silencio} \end{cases} \tag{3.9}$$

Por último, la SNR de cada trama se calcula como:

$$SNR(l) = 10 \cdot log_{10}(P_S(l)) - 10 \cdot log_{10}(P_N(l)) \tag{3.10}$$

Este algoritmo ha sido evaluado y comparado con el sistema original propuesto por Ramirez *et al.* (2004), así como con los definidos en los estándares ITU G.729 (ITU-T, 2007) y ETSI AFE-DSR (ETSI, 2003). G.729 es un estándar de codificación de voz que emplea la VAD para codificar las tramas de silencio de forma eficiente y reducir el ancho de banda requerido. El estándar AFE-DSR define una interfaz de RAH con dos algoritmos VAD, uno utilizado en el sistema de reducción de ruido (AFE-NR) y otro para descartar las tramas de silencio durante el reconocimiento (AFE-FD).

Para las pruebas de precisión se ha utilizado la parte en castellano de la base de datos SpeeCon (Iskra *et al.*, 2002), la cual contiene señales de voz grabadas simultaneamente con cuatro micrófonos, que definen otros tantos canales:

- ***C0***: Micrófono de cercanía (*close-talk*).
- ***C1***: Micrófono de solapa.
- ***C2***: Micrófono cardioide a distancia media (0,5-1 metros).
- ***C3***: Micrófono omnidireccional lejano (2-3 metros).

Cada uno de estos canales representa una SNR diferente, siendo *C0* el más limpio (alrededor de 20 dB) y *C3* el más ruidoso (alrededor de 0 dB). Para el cálculo

de la precisión se han utilizado como referencia etiquetas de actividad vocal corregidas manualmente. A partir de esta referencia y de las etiquetas generadas por los algoritmos se han calculado tres valores:

- **ER0**: El porcentaje de tramas de silencio que se clasifican como voz.
- **ER1**: El porcentaje de tramas con actividad vocal que se clasifican como silencio.
- **TER**: El error total, calculado como el porcentaje total de tramas mal clasificadas.

La descripción completa de las pruebas y los resultados puede encontrarse en las actas de la conferencia LREC (Luengo *et al.*, 2010). Un resumen de estas medidas se muestran en la Tabla 3.1 para cada uno de los sistemas evaluados y cada uno de los canales presentes en la base de datos. Se observa que los algoritmos estándares no son adecuados para aplicaciones de identificación automática de emociones, ya que incluso en situaciones de bajo nivel de ruido (canal *C0*) detectan menos de la mitad de las tramas de silencio. Al tratarse de estándares dedicados a la codificación de voz y al RAH, estos algoritmos están diseñados para minimizar el ER1 (para no perder tramas de voz), mientras que el ER0 no es tan problemático.

El algoritmo LTSE original obtiene también muy buenos resultados en términos de ER1, a la vez que consigue reducir significativamente el valor de ER0, lo cual muestra los beneficios del algoritmo adaptativo. Aún así, el 30 % de las tramas de silencio se clasifican como voz en todos los escenarios. Como ya se ha indicado, esto puede dar lugar a una mala estimación de la distribución de los parámetros. El algoritmo modificado obtiene los valores de ER0 más bajos, entre el 10 % y 20 % dependiendo del nivel de ruido. El ER1 se incrementa considerablemente, pero se mantiene siempre por debajo del 7 %. El error total del sistema también es el más bajo de todos los comparados, con más del 90 % de las tramas correctamente clasificadas en los escenarios *C0*, *C1* y *C2*.

Gracias a las modificaciones realizadas sobre el algoritmo original, se ha logrado un sistema más robusto frente a variaciones de ruido. Se ha reducido el número de tramas de silencio que se clasifican como voz, lo que es muy conveniente para no corromper la distribución de los parámetros calculados. Aunque el número de tramas de voz mal clasificadas se incrementa, no llega el 7 %. Puesto que perder algunas tramas es, en este caso, menos perjudicial, se ha estimado que este algoritmo es especialmente útil en un sistema de identificación automática de emociones.

TABLA 3.1: *Comparación de resultados para los algoritmos considerados en las pruebas. Todos los valores en porcentaje.*

Canal	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C0	56,06	63,88	58,23	38,57	**15,23**
C1	70,23	54,75	55,96	33,04	**8,62**
C2	59,54	52,10	38,10	38,82	**10,25**
C3	70,49	50,10	47,65	34,88	**22,55**

(A) *Tasa de error en tramas de silencio (ER0).*

Canal	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C0	3,63	**0,03**	0,62	0,05	0,78
C1	9,28	**0,23**	1,98	0,49	4,77
C2	18,19	**0,48**	4,83	0,53	6,75
C3	17,22	1,41	8,30	**1,34**	5,04

(B) *Tasa de error en tramas de voz (ER1).*

Canal	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C0	28,98	30,49	28,11	18,68	**7,77**
C1	38,74	26,24	27,73	16,22	**6,63**
C2	38,16	25,09	20,69	19,02	**8,44**
C3	42,94	24,61	27,05	17,54	**13,50**

(C) *Tasa de error total (TER).*

3.1.2. Estimación de la señal glotal

Los parámetros asociados a la calidad de voz (VQ, *Voice Quality*) están relacionados con la forma y las características del pulso glotal. Por tanto, para el cálculo de estos parámetros, es necesario realizar una estimación de la señal glotal. Según el modelo de fuente y filtro (Rabiner y Schafer, 1978), la señal de voz sonora puede modelarse como la salida de un sistema de tres filtros en cascada, tal y como se presenta en la Figura 3.4. $G(z)$ representa la señal periódica generada por las cuerdas vocales, mientras que $V(z)$ y $R(z)$ son el modelo de tracto vocal y el efecto de la radiación labial respectivamente. Si se toma un intervalo de tiempo suficientemente corto como para que $G(z)$, $V(z)$ y $R(z)$ se puedan considerar estacionarios, el espectro de la señal de voz puede expresarse como:

$$S(z) = G(z)V(z)R(z) \tag{3.11}$$

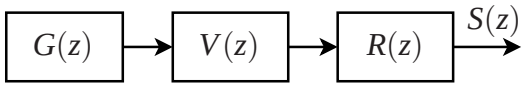


FIGURA 3.4: *Modelo de fuente y filtro de la generación de voz.*

Para el efecto de radiación $R(z)$ suele considerarse suficiente un modelo diferenciador de primer orden:

$$R(z) = 1 - \alpha z^{-1} \tag{3.12}$$

con $0,95 \leq \alpha < 1,0$. Por su parte el filtro del tracto vocal puede modelarse mediante un sistema todo-polos de orden p:

$$V(z) = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-1}} \tag{3.13}$$

donde los coeficientes del filtro a_i pueden aproximarse mediante un análisis de predicción lineal (Rabiner y Schafer, 1978).

Si el efecto de la radiación y el tracto vocal pueden estimarse con suficiente precisión, es posible recuperar la señal glotal mediante filtrado inverso:

$$G(z) = \frac{S(z)}{V(z)R(z)} \tag{3.14}$$

Esto sólo es posible si todas las componentes del modelo ($G(z)$, $V(z)$ y $R(z)$) son linealmente independientes y no interaccionan unas con otras. Desgraciadamente, es sabido que el tracto vocal y las cuerdas vocales son interdependientes, dando como resultado una estimación pobre de la señal glotal. Sin embargo, esta estimación aproximada todavía permite obtener ciertas características generales del pulso glotal, como los intervalos de apertura y cierre, que pueden ser muy útiles a la hora de parametrizar la forma de este pulso (Riegelsberger y Krishnamurthy, 199).

Para la estimación de la señal glotal se ha utilizado el filtrado inverso adaptativo iterativo (IAIF, *Iterative Adaptative Inverse Filtering*) (Alku, 1992), una técnica totalmente automática que proporciona estimaciones aceptables. Este algoritmo estima el filtro del tracto vocal mediante un proceso iterativo, tal y como muestra la Figura 3.5. Primeramente se aplica un filtro paso-alto para eliminar las distorsiones de baja frecuencia capturadas por el micrófono. Posteriormente se aplica un análisis LPC de orden 1 (bloque 2), que proporciona una primera estimación del efecto conjunto del flujo glotal y la radiación labial en el espectro de la señal. Este efecto es eliminado mediante filtrado inverso (bloque 3), y la señal resultante vuelve a analizarse mediante LPC de orden p (bloque 4). Este

segundo análisis proporciona una primera estimación del filtro del tracto vocal. Típicamente suele utilizarse un valor de p entre 8 y 12.

Una vez estimado el efecto del tracto vocal, se elimina de la señal mediante filtrado inverso (bloque 5). La salida de este filtrado inverso es por tanto una aproximación del efecto conjunto de la señal glotal y la radiación de los labios. El efecto de la radiación es eliminado mediante una integración (bloque 6), proporcionando una primera estimación de la señal glotal. Sobre esta señal se aplica un tercer análisis LPC de orden g (bloque 7), que permite calcular la aportación del flujo glotal dentro del espectro de la señal. g está típicamente entre 2 y 4.

Mediante filtrado inverso se elimina el efecto estimado del flujo glotal dentro de la señal (bloque 8), y el efecto de la radiación es cancelado nuevamente por integración (bloque 9). Aplicando un cuarto análisis LPC de orden r (bloque 10) se consigue una aproximación más robusta del filtro del tracto vocal. Por último se aplica un filtrado inverso con esta aproximación (bloque 11), y mediante integración se cancela el efecto de la radiación (bloque 12), dando como resultado final una estimación más robusta del flujo glotal.

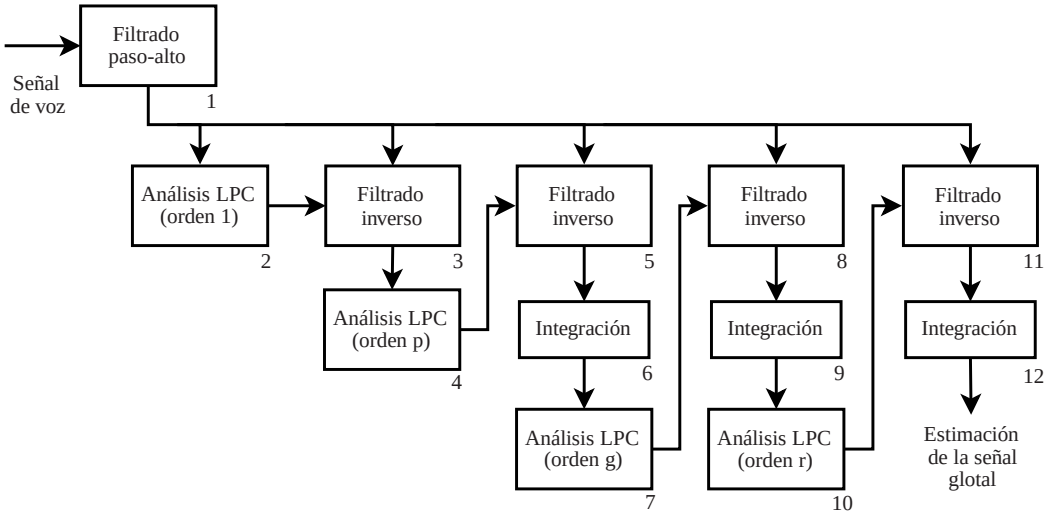
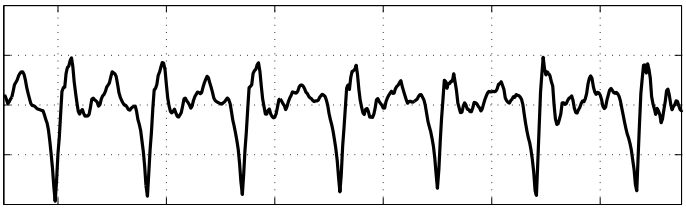


FIGURA 3.5: *Esquema del filtrado inverso IAIF.*

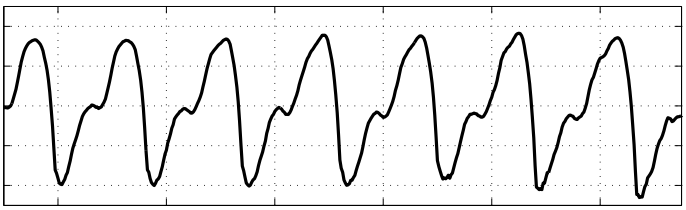
Para el procesado de las señales utilizadas en este trabajo se han fijado los valores $p = 12$, $g = 3$ y $r = 12$ para los análisis LPC. Para el filtrado paso-alto inicial se ha aplicado un filtro FIR con frecuencia de corte 50 Hz. La Figura 3.6 muestra el resultado obtenido al aplicar este algoritmo a un segmento de señal de voz.

La señal glotal estimada de esta forma no es perfecta. Sin embargo, cuando se aplica este método sobre segmentos de la señal suficientemente estables, es posible obtener una aproximación aceptable como para permitir el cálculo de ciertos

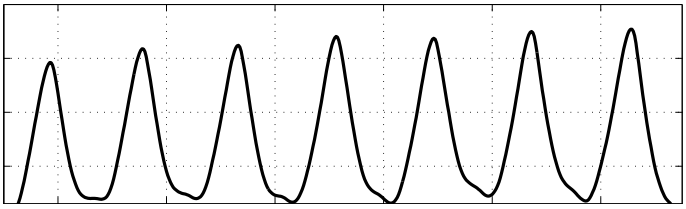
parámetros asociados a la VQ. Por tanto, hay que tener cuidado de calcular estos parámetros sólo en regiones muy estables de la señal.



(A) *Señal de voz.*



(B) *Residuo LPC.*



(C) *Señal glotal estimada.*

FIGURA 3.6: *Resultado de la estimación de la señal glotal mediante IAIF.*

3.1.3. Curva de entonación y decisión sordo-sonoro

La correcta estimación de la curva de entonación es fundamental para poder calcular parámetros prosódicos asociados a la entonación. Existe una gran variedad de algoritmos propuestos para esta tarea (Hess (1983) presenta una revisión de los métodos clásicos), algunos de los cuales son muy populares debido a estar públicamente disponibles o por formar parte de algún paquete de herramientas de procesado de señal (Boersma, 1993; Medan *et al.*, 1991; Talkin, 1995; Xuejing, 2002). Sin embargo, en este trabajo se ha utilizado un algoritmo propio, bautizado como CDP (cepstrum y programación dinámica, *Cepstrum and Dynamic Programming*), y que fue presentado en en la conferencia ICASSP (Luengo *et al.*, 2007). Como su propio nombre indica, CDP se basa en el cálculo de coeficientes cepstrum y en programación dinámica.

Pimeramente la señal de entrada se enventana mediante ventanas de Hamming solapadas. El tamaño de estas ventanas ha de ser suficientemente grande como para abarcar al menos dos períodos de pitch. Para ello se considera que la frecuencia fundamental debe estar dentro del rango [35 Hz–500 Hz], el cual es suficientemente amplio como para contener la verdadera frecuencia fundamental en la mayoría de los casos[1]. Por tanto, se han escogido ventanas de 58 ms. Por cada trama enventanada se calculan los coeficientes cepstrum de la misma:

$$c[k] = \text{IDFT}\{\log(|\text{DFT}\{s[n]\}|)\} \tag{3.15}$$

siendo $s[n]$ las muestras de la trama enventanada y DFT e IDFT las transformadas discretas de Fourier directa e inversa respectivamente. De estos coeficientes cepstrum se eliminan aquellos relativos a frecuencias fuera del rango considerado para la frecuencia fundamental, reteniendo sólo los coeficientes $c[k]$ con $k_{min} < k < k_{max}$:

$$k_{min} = \left\lfloor \frac{F_s}{f_{max}} \right\rfloor \qquad k_{max} = \left\lfloor \frac{F_s}{f_{min}} \right\rfloor \tag{3.16}$$

con $f_{min} = 35$, $f_{max} = 500$ y F_s la frecuencia de muestreo en Hz. Además, para hacer que los valores sean independientes de las variaciones de la intensidad de la señal, los coeficientes se normalizan al valor medio dentro de cada trama:

$$\tilde{c}[k] = \frac{c[k]}{\bar{c}} \qquad \bar{c} = \frac{1}{P} \sum_{k=k_{min}}^{k_{max}} c[k] \tag{3.17}$$

siendo $P = k_{max} - k_{min} + 1$ el número de coeficientes que se han mantenido después de eliminar los que quedan fuera de rango. Este conjunto de P coeficientes normalizados se ordena de mayor a menor y se toman los índices de los primeros M coeficientes como candidatos para la frecuencia fundamental de la trama. A estos M candidatos se les añade uno más, que llamaremos *no-pitch*, representando la opción de que la trama sea sorda. Una vez establecidos los $M+1$ candidatos para cada trama, el algoritmo de programación dinámica se encarga de seleccionar aquellos valores que finalmente formarán la curva de entonación estimada. Además, gracias a que se considera un candidato adicional *no-pitch*, este algoritmo también proporciona la decisión sordo-sonoro (VUV, *Voiced-UnVoiced*).

El algoritmo de programación dinámica busca la secuencia de candidatos que proporciona el mínimo coste total de la curva estimada. Como es tradicional, el

[1] Aunque en ocasiones excepcionales la voz de niños (y sobre todo la voz emocionada) puede superar el umbral de los 500 Hz, incrementar este umbral en exceso provoca un aumento del error en los casos en los que la frecuencia fundamental está en valores más modestos, alrededor de los 100 Hz, lo cual es más habitual.

coste de selección consta de dos componentes: un coste local C^l, asociado a cada candidato e independiente de los candidatos de las tramas adyacentes; y un coste de transición C^t, asociado a los candidatos ya seleccionados. Para definir los valores de estos costes se han utilizado cuatro criterios básicos derivados de las características habituales de una curva de entonación.

1. Cuanto mayor sea el valor de un cepstrum, más probable es que la frecuencia asociada a ese coeficiente sea la frecuencia fundamental.

2. Si el máximo valor para todos los cesptrum de una trama es menor que un cierto umbral, es probable que la trama sea sorda.

3. La curva de entonación varía de forma suave, sin cambios bruscos. Los cambios bruscos suelen aparecer al seleccionar un armónico o subarmónico de la verdadera frecuencia fundamental.

4. Es poco probable que haya transiciones rápidas sordo-sonoro-sordo o sonoro-sordo-sonoro. En caso de que dichas transiciones aparezcan, suelen ser debido a errores en la estimación VUV.

El coste local para el coeficiente k de la trama j se ha definido como:

$$C_j^l[k] = C_j^V[k] + C_j^{thr}[k] \tag{3.18}$$

El primer término de esta expresión está asociado a los candidatos sonoros, y trata de asignar un coste inversamente proporcional al valor del cepstrum. Es decir, cuanto mayor sea el valor del cepstrum, menor será el coste (criterio 1). Para el candidato *no-pitch* este coste es cero:

$$C_j^V[k] = \begin{cases} -W_V \log\left(\tilde{c}_j[k]\right) & k = 1 \ldots M \\ 0 & k = M+1 \end{cases} \tag{3.19}$$

El segundo término de (3.18) está asociado al criterio 2. Para los candidatos sonoros es cero si el valor del coeficiente cepstrum supera un umbral T, es decir, si el valor del cepstrum es suficientemente grande como para suponer que la trama es sonora. En el caso del candidato *no-pitch*, el coste es cero sólo si *ningún* candidato sonoro supera este umbral. En cualquier otro caso el coste tendrá un valor fijo W_{thr}. En lugar de suponer que la trama es sorda si ningún cepstrum supera el umbral, esta implementación delega en el algoritmo de programación dinámica para tomar la decisión VUV, pudiendo romper este umbral (con un coste W_{thr}) si lo considera oportuno.

$$C_j^{thr}[k] = \begin{cases} 0 & k \in [1,M] \quad \wedge \quad \tilde{c}_j[k] > T \\ 0 & k = M+1 \quad \wedge \quad \tilde{c}_j[k] < T \ \forall k = 1 \ldots M \\ W_{thr} & \text{otro} \end{cases} \tag{3.20}$$

El coste de transición se define en función de los criterios 3 y 4: se favorecen las transiciones entre candidatos sonoros si sus frecuencias correspondientes son próximas (evitando saltos bruscos en la curva), mientras que las transiciones sordo-sonoro y sonoro-sordo tienen una penalización constante para evitar transiciones rápidas:

$$C_{j,j-1}^{t} = \begin{cases} W_{cont} \left| \log\left(\frac{f_j}{f_{j-1}}\right) \right| & \text{transición V-V} \\ W_{VUV} & \text{transición V-U o U-V} \\ 0 & \text{transición U-U} \end{cases} \tag{3.21}$$

Los valores de los parámetros M, T, W_V, W_{thr}, W_{cont} y W_{VUV} han sido estimados empíricamente de forma que el sistema proporcione resultados aceptables en diferentes bases de datos y entornos de ruido. Estos valores vienen reflejados en la Tabla 3.2. Para comprobar el correcto funcionamiento de este algoritmo, se han llevado a cabo una serie de pruebas, comparando los resultados obtenidos con algunos sistemas ampliamente utilizados:

- **AM**: Algoritmo del programa Praat[2], basado en el cálculo preciso de la autocorrelación (Boersma, 1993).
- **RAPT**: Algoritmo implementado en el programa WaveSurfer/ESPS[3], basado en correlación cruzada y programación dinámica.
- **SHR**: Algoritmo basado en la relación subarmónico a armónico, tal y como lo describe Sun (2000).
- **SRPD**: Determinador de pitch de alta resolución (*Super Resolution Pitch Determinator*)(Medan *et al.*, 1991), según la implementación de la biblioteca de herramientas del habla de Edimburgo[4].

Las pruebas se han realizado sobre la base de datos SpeeCon anteriormente citada (Iskra *et al.*, 2002). Como referencia se han tomado curvas de entonación ajustadas manualmente. La precisión de los sistemas se ha evaluado en función de las siguientes medidas:

[2] www.praat.org
[3] www.speech.kth.se/wavesurfer
[4] www.cstr.ed.ac.uk

TABLA 3.2: *Valores de los parámetros del algoritmo CDP.*

Parámetro	Valor
M	5
W_V	2
W_{thr}	2
W_{cont}	200
W_{VUV}	30
T	4

- La tasa de errores de clasificación VUV, calculada como el porcentaje de tramas mal clasificadas.

- La tasa de errores severos, definida como el porcentaje de tramas con un error superior al 20 % en el cálculo del valor de F_0. Estos errores suelen ocurrir cuando el algoritmo detecta un armónico o subarmónico en lugar del verdadero valor de F_0. Para este cálculo sólo se utilizan las tramas sonoras que han sido detectadas como tales.

- La raíz del error cuadrático medio (RMSE, *Root Mean Square Error*) de las muestras de F_0. Para este cálculo no se consideran las tramas con errores severos. Por lo tanto, este valor mide la precisión fina de la curva estimada.

Los resultados de estas pruebas se detallan en el artículo original que describe el algoritmo (Luengo *et al.*, 2007). La Tabla 3.3 presenta un resumen de estos resultados. Para niveles de ruido reducidos (canal *C0*), los métodos basados en correlación son los que menos errores cometen en la clasificación VUV, mientras que el algoritmo propuesto está entre los peores. Sin embargo, estos métodos sufren un brusco incremento de los errores en los entornos ruidosos (canales *C1*, *C2* y *C3*). Por el contrario, el algoritmo CDP mantiene mejor sus resultados, situándose en primera posición para estos canales.

En cuanto a los errores severos, el sistema propuesto supera ampliamente a los demás, manteniendo los niveles de error significativamente bajos incluso en entornos de muy baja SNR (canal *C3*). Analizando los errores finos, los algoritmos RAPT y SRPD obtienen los mejores valores de RMSE, seguidos de cerca por el CDP.

Los resultados de estas pruebas certifican que CDP es un algoritmo de extracción de pitch robusto, sobre todo en entornos de ruido, superando en muchos casos a los demás sistemas evaluados. Si bien el valor de RMSE es ligeramente superior

al obtenido por los sistemas RAPT y SRPD, comete muchos menos errores severos. En comparación con los demás algoritmos, la decisión VUV es algo pobre en entornos libres de ruido, pero resulta ser la más precisa cuando se trabaja con señales de baja SNR.

TABLA 3.3: *Comparación de resultados para los algoritmos considerados en las pruebas de detección de entonación.*

Canal	CDP	AM	RAPT	SHR	SRPD
C0	18,83	**12,20**	16,47	24,25	18,13
C1	**28,64**	32,56	36,22	42,06	45,18
C2	**35,32**	37,12	36,15	41,44	53,62
C3	**43,20**	52,93	54,65	63,24	72,57

(A) *Tasa de errores de clasificación VUV (%).*

Canal	CDP	AM	RAPT	SHR	SRPD
C0	**0,83**	5,88	2,60	4,70	3,11
C1	**0,72**	15,57	2,68	4,47	4,04
C2	**0,83**	13,93	3,50	7,13	4,42
C3	**1,19**	17,48	5,06	6,53	5,60

(B) *Tasa de errores severos (%).*

Canal	CDP	AM	RAPT	SHR	SRPD
C0	5,21	11,8	4,30	7,00	**4,14**
C1	4,91	12,83	4,23	6,91	**4,17**
C2	5,48	12,21	**4,19**	6,58	4,25
C3	6,47	13,96	**5,53**	7,15	6,20

(C) *RMSE (Hz).*

3.1.4. Marcas a período de pitch

Además de la curva de entonación, también es importante disponer de marcas a período de pitch en la señal, ya que estas marcas son necesarias para el cálculo de parámetros asociados a la calidad de voz.

Una vez estimadas la curva de entonación, la decisión VUV y la señal glotal, las marcas a período de pitch se han colocado en los picos negativos del resíduo LPC (obtenido mediante el algoritmo IAIF). Estos picos negativos se han detec-

tado con un sencillo algoritmo de selección de mínimos (*peak-picking*), utilizando la curva de entonación estimada como guía para detectar la posición del siguiente mínimo.

3.1.5. Detección de la posición de las vocales

Conocer los segmentos de voz correspondientes a las vocales puede ser muy útil a la hora de realizar la parametrización. Por un lado, las vocales representan segmentos muy estables de la señal, lo que hace de ellas un lugar apropiado a la hora de calcular parámetros relacionados con la señal glotal. Por otro lado, el número de vocales por unidad de tiempo puede dar una medida aproximada del ritmo del habla (Pfau y Ruske, 1998). Varios de los parámetros utilizados en este trabajo toman como referencia la posición de las vocales (ver sección 3.2).

Puesto que el objetivo de esta tesis es analizar el comportamiento de parámetros calculados de forma automática, también la posición de las vocales se ha estimado automáticamente. Para ello se ha utilizado un sistema de reconocimiento basado en HMM de agrupaciones fonéticas y una gramática libre en forma de bucle infinito de modelos concatenados (Luengo *et al.*, 2009b). La Figura 3.7 presenta el esquema del detector. Como puede verse, las señales a procesar son primeramente analizadas por el sistema VAD descrito en la sección 3.1.1, para eliminar los segmentos de silencio. Sólo las tramas clasificadas como voz son procesadas por el detector de vocales.

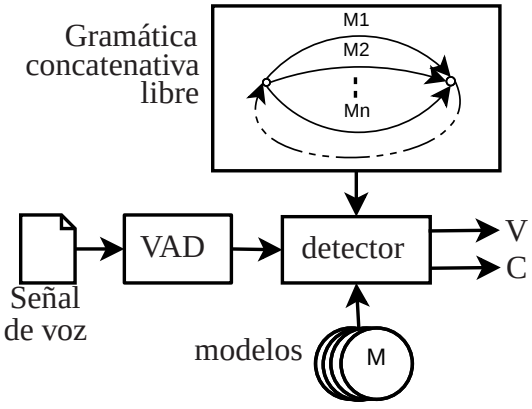


FIGURA 3.7: *Esquema del detector de vocales.*

Este detector fue originalmente entrenado para detectar las vocales en grabaciones en euskera. Los modelos fueron entrenados utilizando la base de datos SpeechDat–EU (Hernáez *et al.*, 2003), que contiene grabaciones en euskera realizadas a 1050 locutores a través de la red de telefonía fija. Este entrenamiento se

realizó utilizando las herramientas de HTK (Young *et al.*, 2000) y siguiendo el sistema de reconocimiento de referencia RefRec (Lindberg *et al.*, 2000). El número de componentes gaussianas de los modelos se seleccionó empíricamente, estableciéndose en 1024. En cuanto a la parametrización, se emplearon 12 parámetros MFCC junto con sus primeras y segundas diferencias y con normalización de la media cepstral (CMS, *Cepstral Mean Subtraction*).

La agrupación de fonemas es clave en el comportamiento de este sistema. Si cada fonema se modelara por separado, la tasa de error del sistema sería muy grande, tanto en la secuencia de los fonemas detectados como en la precisión temporal de las marcas estimadas. Esta tasa de error está relacionada con la gran cantidad de alternativas que tiene el sistema para decidir el fonema correspondiente a cada instante de tiempo. Recordemos que se trata de una gramática sin restricciones, donde cualquier modelo puede seguir a cualquier otro. Se implementó de esta manera para minimizar la dependencia con el idioma, ya que imponer restricciones en las secuencias de fonemas posibles impediría el uso del sistema en otros idiomas. Para reducir la complejidad se llevó a cabo una agrupación ciega de los fonemas según su similitud acústica. Por cada grupo fonético resultante se entrenó un único modelo, de forma que el sistema no ha de escoger entre todos los fonemas existentes, sino sólo a qué grupo pertenece. Al reducir el número de alternativas, también se reduce la tasa de error.

La agrupación de fonemas se realizó mediante árboles de regresión. La Figura 3.8 muestra el dendrograma con todos los fonemas del euskera (expresados en código SAMPA[5]) y el clustering resultante. El punto de poda del árbol está representado por la línea discontinua, y fue seleccionado teniendo en cuenta tanto la similitud acústica de los grupos como el número de ejemplos de entrenamiento disponibles para cada grupo, con el objetivo de asegurar el entrenamiento de modelos robustos. Se puede observar que los grupos de fonemas resultantes se corresponden aproximadamente con los diferentes modos de articulación, de tal forma que en general se puede identificar el grupo de fonemas fricativos y africados o el grupo de nasales. El grupo denominado *L y similares* contiene fonemas que en principio no comparten modo de articulación (/L/ es líquido, /jj/ es fricativo y /gj/ es africado), pero sin embargo no es un grupo excesivamente heterogéneo. Todos ellos tienen un sonido bastante similar, hasta tal punto que muchas personas no distinguen entre ellos y los pronuncian igual. Se cree que la razón de que hayan quedado en el mismo grupo es precisamente que muchos locutores de la base de datos no diferenciaban correctamente estos fonemas al pronunciarlos.

Aunque el detector fonético fue diseñado para trabajar con grabaciones en euskera, gracias a la agrupación fonética se consigue que también sea capaz de obtener resultados aceptables en otros idiomas. Si cada fonema se modelara por

[5] http://aholab.ehu.es/sampa_basque.htm

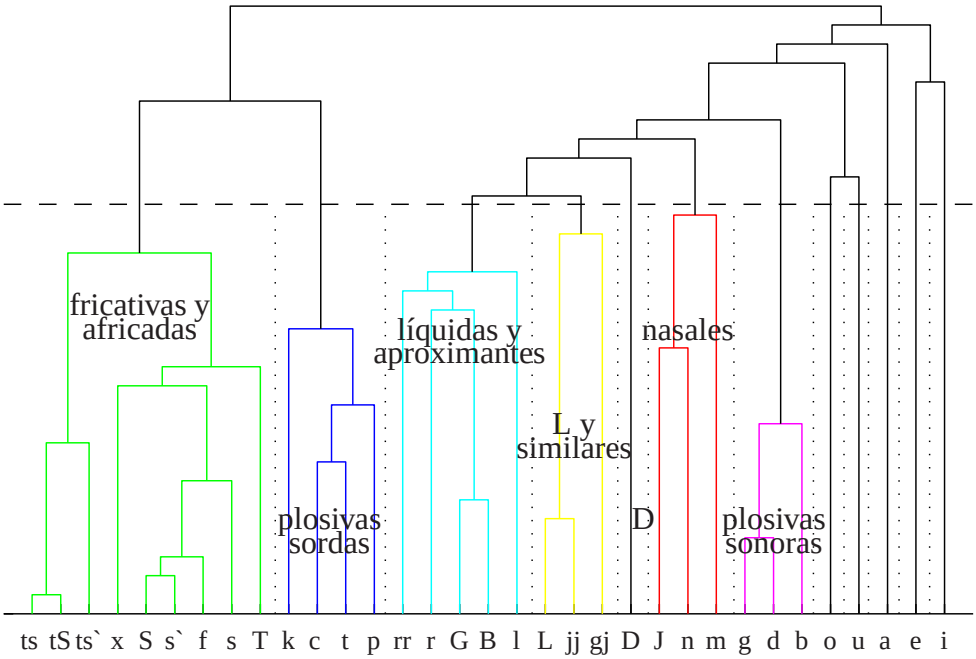


FIGURA 3.8: *Dendrograma de la salida del clustering de fonemas. Cada grupo de fonemas está representado por un color diferente, excepto los grupos que constan de un único fonema, que están representados en negro. Los fonemas están expresados mediante código SAMPA.*

separado, el sistema estaría excesivamente especializado, y no podría utilizarse en un idioma que tuviera un conjunto de fonemas diferente. Realizando la agrupación fonética se consigue que cada modelo recoja fonemas de características similares, aumentando la robustez, a la vez que se asegura que los modelos sean capaces de detectar fonemas no conocidos, asignándolos a su grupo correspondiente. De esta forma la precisión obtenida en idiomas diferentes al de entrenamiento no sufre un descenso tan acusado.

Se han realizado varios experimentos para comprobar la eficacia del sistema propuesto, tanto en euskera (el idioma original para el que fue entrenado el sistema) como en alemán, un idioma con un sistema vocálico muy diferente. Mientras que el euskera contiene 5 vocales /a,e,i,o,u/, el alemán contiene 9 vocales cortas y 7 vocales largas, siendo la duración de las vocales largas aproximadamente el doble que el de las cortas. De estas vocales, 4 cortas y 2 largas no tienen correspondencia en el euskera, con lo que la detección de las mismas puede ser complicada. Debido a esto, los resultados de detección de vocales en este idioma son muy significativos. Para las pruebas se utilizaron las bases de datos *AhoEmo2*

(Saratxaga *et al.*, 2006) (en euskera) y *Berlin* (Burkhardt *et al.*, 2005) (en alemán). Ambas disponen de una segmentación fonética de referencia.

Las principales medidas para determinar el funcionamiento del sistema han sido la precisión de detección y la precisión temporal de las marcas. La precisión de detección se calcula teniendo en cuenta tanto los errores de inserción como de omisión en las vocales detectadas, y se define como:

$$Acc = \frac{N_{ref} - E_{ins} - E_{omi}}{N_{ref}} \times 100 \tag{3.22}$$

dónde N_{ref} es el número total de vocales en el etiquetado de referencia, E_{ins} es el número de errores de inserción y E_{omi} es el número de errores de omisión.

Puesto que el objetivo de esta detección de vocales es poder utilizar los segmentos vocálicos para el cálculo de parámetros relativos a la voz, de poco sirve detectar una vocal si no podemos determinar sus fronteras de forma precisa. Por ello también es importante la precisión temporal de las marcas creadas. Esta precisión temporal se ha medido como el porcentaje de marcas automáticas con un error inferior a 20 ms. Se trata de la medida más utilizada para la precisión temporal de las marcas, lo que permite comparar los resultados con otros sistemas. Un valor superior a 80 % suele considera bueno (Hosom, 2000).

La Tabla 3.4 resume los principales resultados de las pruebas. Se comparan los resultados obtenidos con el sistema propuesto y sin agrupación fonética, es decir, modelando cada fonema por separado. Como se puede observar en los resultados, la agrupación fonética implementada mejora la precisión de la detección de las vocales para el idioma de entrenamiento (euskera). Además, esta agrupación hace posible aplicar este método a otros idiomas con una estructura fonética diferente, lo cual sería imposible si se modelara cada fonema por separado. En el artículo original de Luengo *et al.* (2009b) puede encontrarse la descripción completa de las pruebas y todos los resultados.

3.2. Definición de los parámetros

Los parámetros estudiados se han dividido en dos grandes grupos: los parámetros segmentales (calculados para cada trama) y los parámetros supra-segmentales (calculados sobre intervalos de integración largos). Esta división obedece al hecho de que cada uno de los dos grupos tiene diferentes características que afectan a la elección del clasificador utilizado (ver sección 2.3).

TABLA 3.4: *Resultados de las pruebas de detección de vocales. Todos los valores en porcentaje.*

Base de datos	Omisiones	Inserciones	Precisión	Temp. <20ms
		Con la agrupación propuesta		
AhoEmo2	7,79	6,58	85,71	86,10
Berlin	9,70	18,89	69,27	75,91
		Sin agrupación fonética		
AhoEmo2	18,32	5,84	76,62	84,56
Berlin	73,87	71,20	-38,46	27,81

3.2.1. Parámetros segmentales

Los parámetros segmentales permiten analizar la evolución temporal de sus valores, ya que se calculan cada corto plazo. Todos los parámetros a nivel de trama se han calculado cada 10 ms utilizando un enventanado de Hamming de 25 ms con un solapamiento de 15 ms entre tramas consecutivas.

Espectrales

Debido a la influencia del campo del RAH, la mayoría de los sistemas de identificación de emociones que utilizan información espectral a nivel de trama utilizan parámetros MFCC o LPCC. Sin embargo, según el trabajo de Nwe *et al.* (2003), los LFPC tienen una mayor capacidad para discriminar emociones. Por tanto, se ha optado por utilizar LFPC para la caracterización de la envolvente espectral.

Los LFPC miden la potencia en bandas de energía dispuestas según una escala mel. Su cálculo es similar al de los MFCC pero sin la transformada coseno discreta final. La Figura 3.9 muestra el diagrama de bloques correspondiente.

La señal es primeramente preenfatizada (bloque 1) para compensar el efecto de la radiación en los labios. Como coeficiente de preénfasis se ha utilizado un valor $\alpha = 0,95$. Una vez enventanadas las tramas (bloque 2), se obtiene el espectro de potencia (bloques 3 y 4) y se calcula la potencia de cada banda según un banco de filtros triangulares en escala mel (bloque 5). Por último, se convierten estos valores a escala logarítmica (bloque 6). En este caso se han utilizado 18 bandas de frecuencia, lo que ha proporcionado 18 parámetros por cada trama.

Para reducir el efecto de las variaciones en la intensidad, se ha aplicado una normalización de media (bloque 7): una vez parametrizada la señal y descartadas las tramas de silencio detectadas por el VAD, se ha calculado el valor medio de

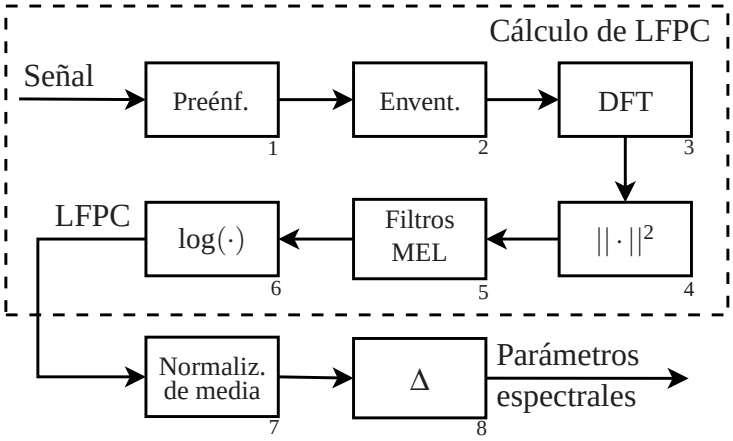


FIGURA 3.9: *Diagrama de bloques de la parametrización LFPC.*

los parámetros de las tramas de voz y se ha sustraído este valor medio a todos los vectores de la señal. Por último, se han añadido las primeras y segundas derivadas a los vectores de LFPC (bloque 8) para retener parte de la información dinámica. Las derivadas se han estimado como la pendiente de la recta de regresión para un entorno de dos tramas alrededor de la actual. Formalmente, la derivada de un parámetro P en la trama i se ha estimado como:

$$\Delta P[i] = \frac{\sum_{k=1}^{2} (P[i+k] - P[i-k])}{\sum_{k=1}^{2} 2k^2} \tag{3.23}$$

Para la segunda derivada se ha aplicado el mismo procedimiento, tomando como primitiva la curva de la primera derivada, ya calculada. Los 18 LFPC junto con sus primeras y segundas derivadas proporcionan un total de $18 \times 3 = 54$ parámetros espectrales a corto plazo.

Primitivas de prosodia

Con el término *primitivas de prosodia* denominamos a las curvas de entonación e intensidad, ya que los parámetros prosódicos (supra-segmentales) se calculan a partir de ellas. La Figura 3.10 presenta el diagrama del cálculo de estos parámetros. Puesto que el valor de F_0 no está definido en las tramas sordas, las primitivas de prosodia se han separado en dos flujos de parámetros, uno para las tramas sonoras y otro para las sordas. El flujo de las tramas sonoras está formado por vectores de seis parámetros (F_0 y potencia, junto con sus primeras y segun-

das diferencias) mientras que el de las tramas sordas contiene vectores de sólo tres parámetros (potencia y sus primeras y segundas diferencias). Ambos flujos se han tratado como parametrizaciones independientes a efectos del análisis de parámetros llevado a cabo.

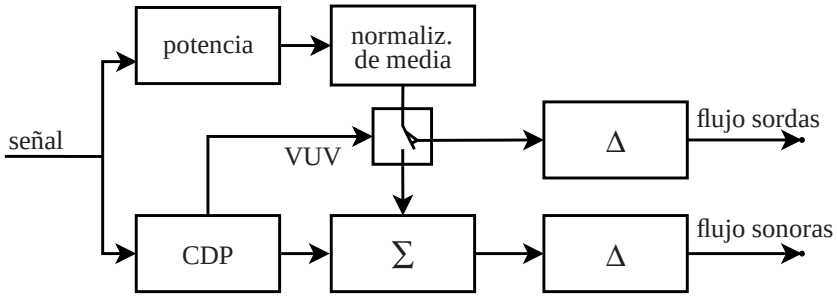


FIGURA 3.10: *Diagrama de la parametrización de primitivas de prosodia.*

Los valores de potencia se han normalizado a la potencia media de la señal para suavizar los efectos provocados por cambios en la distancia al micrófono. Al igual que en el caso de los LFPC, este valor medio se ha calculado descartando las tramas de silencio detectadas por el VAD. El valor de F_0 es el obtenido mediante el algoritmo CDP, mientras que las diferencias se han calculado utilizando la expresión (3.23).

3.2.2. Parámetros supra-segmentales

Los parámetros supra-segmentales son aquellos que recogen información a largo plazo. A la hora de utilizar este tipo de parámetros es importante escoger adecuadamete el intervalo de tiempo sobre el que se calculan dichos parámetros o *intervalo de integración*. Puesto que se obtiene un único vector de parámetros por cada intervalo, un tiempo de integración excesivamente largo proporciona pocos datos y poca resolución temporal. Por el contrario, un intervalo excesivamente corto resulta insuficiente para capturar la evolución de la curva y su estructura a largo plazo. Además, la definición del intervalo de integración no debe realizarse en términos de tiempo absoluto (por ejemplo, "*5 segundos*"), ya que la locución se cortaría por puntos arbitrarios, posiblemente poco adecuados, como por ejemplo en mitad de una frase o una palabra. Estos puntos de corte deben buscarse en función de la propia estructura y contenidos de la locución.

A falta de una definición consensuada de cuál debería ser el intervalo de integración, frecuentemente se suele utilizar la duración de toda la grabación, es decir, se obtiene un único vector de parámetros por cada grabación. Cuando las grabaciones constan de una única frase, esta aproximación puede resultar aceptable. Sin

embargo, si se consideran aplicaciones en las que cada locución puede tener varias frases, se pierde precisión temporal. Puede que sólo una parte de la locución contenga voz emocionada, con lo que puede ocurrir que la emoción detectada se asigne erróneamente a toda la grabación, o que no se detecte emoción alguna, al estar las características mezcladas con voz no emocionada.

Para solventar este problema se ha seleccionado como tiempo de integración el intervalo entre dos silencios consecutivos. A grandes rasgos, se espera que los silencios coincidan con pausas lingüísticas en el mensaje (equivalentes a puntos o comas en la transcripción), con lo que esta aproximación es muy similar a utilizar una frase completa. Sin embargo, ofrece mayor flexibilidad en el caso de locuciones más largas, ya que permite obtener diferentes segmentos y clasificar cada uno por separado, proporcionando mayor resolución temporal en la detección de emociones.

Para la detección de los silencios se ha aplicado el algoritmo VAD descrito en la sección 3.1.1. Por cada segmento detectado como voz por el algoritmo, se ha calculado un vector de parámetros supra-segmentales, con los valores descritos a continuación.

Espectrales

Para la caracterización supra-segmental del espectro se han calculado estadísticos a largo plazo de los valores LFPC. Por cada coeficiente LFPC y sus primeras y segundas derivadas se han estimado los estadísticos reflejados en la Tabla 3.5. Puesto que el intervalo de integración utilizado es relativamente largo, se dispone de suficientes muestras como para calcular de forma robusta los estadísticos de mayor orden como el sesgo o la kurtosis.

Teniendo en cuenta que se dispone de 18 parámetros LFPC y se calculan 6 estadísticos por cada uno de ellos y sus derivadas, se han utilizado un total de $18 \times 3 \times 6 = 324$ parámetros espectrales supra-segmentales.

Prosódicos

Los parámetros prosódicos supra-segmentales se han dividido en cinco categorías, en función de la parte de la información prosódica que tratan de recoger. En total se han definido 54 parámetros.

Parámetros de entonación

Los *parámetros de entonación* recogen la información de F_0 a lo largo de todo el intervalo de integración. Se han definido como los estadísticos presentados en la Tabla 3.5 aplicados a las curvas de F_0 y sus primeras y segundas derivadas. Por tanto este grupo define $3 \times 6 = 18$ parámetros.

TABLA 3.5: *Estadísticos utilizados para el cálculo de los parámetros suprasegmentales de LFPC, entonación e intensidad, junto con el símbolo utilizado a lo largo de este documento. Por ejemplo,* $\mathrm{E}(\Delta LFPC_{10})$ *representa el valor medio de la primera derivada del décimo coeficiente LFPC.*

Parámetro	Símbolo
Valor medio	$\mathrm{E}(\cdot)$
Varianza	$\sigma^2(\cdot)$
Mínimo	$\min(\cdot)$
Rango	$\mathrm{R}(\cdot)$
Sesgo	$\mathrm{Sk}(\cdot)$
Kurtosis	$\mathrm{K}(\cdot)$

Para el cálculo de estos estadísticos sólo se utilizaron las tramas detectadas como sonoras por el algoritmo CDP, ya que el valor de F_0 no está definido para las tramas sordas.

Parámetros de intensidad

Los *parámetros de intensidad* describen la curva de potencia de la señal a lo largo de todo el intervalo de integración. Al igual que en el caso de la entonación, se han definido como los estadísticos de la Tabla 3.5 aplicados a las curvas de potencia y sus primeras y segundas derivadas. También en este caso se definen $3 \times 6 = 18$ parámetros.

Parámetros de ritmo

Los *parámetros de ritmo* describen la señal en términos de la velocidad del habla. Desde el punto de vista lingüístico, la medida del ritmo depende de la naturaleza del idioma hablado: silábico, acentual o basado en moras. En la práctica se suele considerar que una medida del ritmo basada en la velocidad silábica es un buen compromiso, con resultados aceptables en todos los idiomas (Pellegrino y Andre-Obrecht, 2000). Sin embargo, la segmentación silábica de una señal de voz es un proceso también dependiente del idioma. Aunque existen algunas características acústicas universales que permiten desarrollar silabificadores automáticos, la precisión de estos sistemas depende del idioma sobre el que se utilizan (Pellegrino y Andre-Obrecht, 2000).

Algunos autores simplifican el problema seleccionando como medida del ritmo la inversa de la duración media de los segmentos sonoros de la señal (Banse y Scherer, 1996; Tato *et al.*, 2002). Aunque se tata de una medida poco

precisa para el cálculo del ritmo en sentido estricto, sí está asociada a la velocidad del habla, y es mucho más fácil de calcular que una segmentación silábica.

Otra alternativa a la segmentación silábica es utilizar el número de vocales por unidad de tiempo (Pfau y Ruske, 1998). Puesto que en la mayoría de los casos una sílaba contiene una única vocal o un diptongo, el número de sílabas está fuertemente correlado con el número de vocales, por lo que se espera obtener una medida del ritmo más adecuada.

Teniendo en cuenta que ya se dispone de una detección automática de vocales (sección 3.1.5), se ha optado por definir los parámetros de ritmo en función de las vocales detectadas. Las dos medidas utilizadas se indican en la Tabla 3.6.

TABLA 3.6: *Estadísticos utilizados para la caracterización del ritmo.*

Parámetro	**Símbolo**
Duración media de las vocales	$\mathrm{E}(Vdur)$
Varianza de la duración de las vocales	$\sigma^2(Vdur)$

Parámetros de regresión

Con el objetivo de combinar intervalos de integración largos (de duración aproximada de una frase) y cortos (de duración comparable a una sílaba, algunos autores consideran utilizar simultáneamente estadísticos globales sobre la frase completa y tendencias locales sobre segmentos cortos (Ringeval y Chetouani, 2008). La idea es estimar una curva de regresión sobre estos intervalos cortos y calcular uno o varios estadísticos en función de los parámetros de estas regresiones.

Puesto que ya se dispone de la detección automática de vocales, se han utilizado los segmentos vocálicos detectados como segmentos de corta duración. Por cada vocal detectada se ha calculado la regresión lineal de las curvas de F_0 y potencia dentro de los límites de la vocal. A partir de los valores absolutos de la pendiente de estas regresiones se han estimado los 6 parámetros descritos en la Tabla 3.7.

Parámetros de fin de frase

Los valores prosódicos relativos al final de una frase pueden proporcionar información adicional, ya que muchas veces la emoción tiene un efecto especial en esta región. Por ejemplo, un incremento de F_0 con respecto al resto de la frase puede significar sorpresa, mientras que una intensidad notablemente inferior se

TABLA 3.7: *Parámetros derivados de la regresión lineal de F_0 y potencia en los segmentos vocálicos.*

Parámetro	**Símbolo**
Media de la pendiente de F_0	$\mathrm{E}(F_0sl)$
Varianza de la pendiente de F_0	$\sigma^2(F_0sl)$
Máximo de la pendiente de F_0	$\max(F_0sl)$
Media de la pendiente de potencia	$\mathrm{E}(Psl)$
Varianza de la pendiente de potencia	$\sigma^2(Psl)$
Máximo de la pendiente de potencia	$\max(Psl)$

asocia a emociones de poca tensión, como el aburrimiento o la tristeza. Para poder capturar estos efectos, se ha tomado la última vocal detectada para el intervalo de integración y se han calculado los 10 parámetros definidos en la Tabla 3.8. Puesto que las características asociadas al final de la frase pueden ser significativas en relación al resto de la frase, y no en términos absolutos, se incluyen medidas normalizadas, definidas como el valor no normalizado dividido por el valor medio para todas las vocales del intervalo de integración (por ejemplo: $NLvF0sl = \frac{LvF0sl}{\mathrm{E}(F0sl)}$).

TABLA 3.8: *Parámetros utilizados para capturar la tendencia de la prosodia al final de la frase.*

Parámetro	**Símbolo**
Pendiente de F_0 en la última vocal	LvF_0sl
Media de F_0 en la última vocal	LvF_0cn
Pendiente de potencia la última vocal	$LvPsl$
Media de potencia en la última vocal	$LvPcn$
Duración de la última vocal	$LvVdur$
Pendiente normalizada de F_0 en la última vocal	$NLvF_0sl$
Media normalizada de F_0 en la última vocal	$NLvF_0cn$
Pendiente normalizada de potencia la última vocal	$NLvPsl$
Media normalizada de potencia en la última vocal	$NLvPcn$
Duración normalizada de la última vocal	$NLvVdur$

Calidad de voz

Los parámetros de calidad de voz se calculan a partir de la señal glotal o del residuo LPC, obtenidos mediante el filtrado inverso de la voz. Este tipo de parámetros no es muy utilizado debido a la dificultad de obtener una correcta estimación de la señal glotal. El proceso de filtrado inverso no proporciona un buen resultado a no ser que se aplique sobre regiones muy estables de la señal (Gobl y Chasaide, 2003). Puesto que se considera que las vocales son precisamente regiones muy estables en la voz, se ha decidido calcular estos parámetros sólo en las vocales detectadas por el algoritmo presentado en la sección 3.1.5. En cada una de estas vocales se ha aplicado el algoritmo de filtrado inverso IAIF descrito en la sección 3.1.2, y a partir de la señal glotal obtenida se han calculado los parámetros descritos en la Tabla 3.9, dando como resultado un vector de cinco parámetros por cada vocal detectada. Finalmente, se han promediado los vectores correspondientes a las vocales detectadas dentro de un mismo intervalo de integración, dando lugar a un único vector de parámetros para cada intervalo.

TABLA 3.9: *Parámetros relativos a la calidad de voz.*

Parámetro	**Símbolo**
Jitter	*Jit*
Shimmer	*Shm*
Coeficiente de amplitud normalizado	*NAQ*
Pendiente espectral	*TLT*
Equilibrio espectral	*SB*

Jitter

El *jitter* se define como las microvariaciones de la curva de entonación. Para su cálculo se han utilizado las marcas a período de pitch descritas en la sección 3.1.4. Este cálculo se basa en la definición del parámetro ppq5 (cociente de perturbación de período de cinco puntos, *Five-point Period Perturbation Quotient*) de Praat[6]. Este parámetro se define como el valor medio de la diferencia absoluta entre un período y la media de un entorno de 5 períodos centrado en él, normalizado por el período medio de la señal. La expresión matemática correspondiente es:

[6] www.praat.org

$$ppq5 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left|T_i - \frac{1}{2K+1}\sum_{k=-K}^{K} T_{i+k}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{3.24}$$

siendo N el número de períodos de F_0 dentro de la vocal, $K = 2$ y T_i el período i-ésimo. Esta expresión tiene la desventaja de utilizar el período medio de toda la señal como término de normalización, con lo que no tiene en cuenta la variación natural del período debido a la prosodia. Por ello se ha utilizado esta otra expresión, bastante similar:

$$\widehat{ppq5} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|T_i - \frac{1}{2K+1}\sum_{k=-K}^{K} T_{i+k}\right|}{\frac{1}{2K+1}\sum_{k=-K}^{K} T_{i+k}} \tag{3.25}$$

En este caso la diferencia absoluta se normaliza al período medio del mismo entorno de 5 períodos.

Shimmer

El *shimmer* mide las variaciones de intensidad entre períodos consecutivos. El cálculo del shimmer se basa en la definición del parámetro apq5 (cociente de perturbación de amplitud de cinco puntos, *five-point Amplitude Perturbation Quotient*) de Praat, que se define como el valor medio de la diferencia absoluta entre la amplitud pico-a-pico de un período y la media de un entorno de 5 períodos centrado en él, normalizado por la amplitud pico-a-pico media de la frase. Dicho de otra forma:

$$apq5 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left|A_i - \frac{1}{2K+1}\sum_{k=-K}^{K} A_{i+k}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{3.26}$$

siendo N el número de períodos dentro de la vocal, $K = 2$ y A_i la amplitud pico-a-pico del período i-ésimo. Esta expresión es equivalente a (3.24) para el jitter, y está sometida a los mismos problemas, al normalizarse también al valor medio de toda la frase. Por lo tanto, se ha utilizado la siguiente expresión modificada:

$$\widehat{apq5} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|A_i - \frac{1}{2K+1}\sum_{k=-K}^{K} A_{i+k}\right|}{\frac{1}{2K+1}\sum_{k=-K}^{K} A_{i+k}} \tag{3.27}$$

Para el cálculo de las amplitudes A_i se localiza el período i-ésimo utilizando las marcas a período de pitch y se busca el valor máximo y mínimo de la señal de voz dentro de este período. A_i es la diferencia entre estos dos valores.

Cociente de amplitud normalizada (NAQ)

El cociente de amplitud normalizada (NAQ, *Normalized Amplitude Quotient*) es un valor asociado a la duración de la fase de cierre de las cuerdas vocales, y está estrechamente relacionado con el cociente de cierre (CQ, *Closing Quotient*) (Bäckström *et al.*, 2002). El CQ se define como la relación entre el tiempo de cierre de las cuerdas vocales y el período total de pitch. En referencia a la Figura 3.11:

$$CQ = \frac{T_{close}}{T_0} \tag{3.28}$$

Es difícil medir este valor de forma precisa, pues no es sencillo detectar los instantes de apertura y cierre de la glotis con una precisión suficiente. Menos aún si se trata de una señal natural procesada mediante filtrado inverso sin supervisión, ya que presentará bastante ruido en la estimación de la señal glotal. Sin embargo, es posible hallar una medida equivalente basada en amplitudes en lugar de instantes de tiempo.

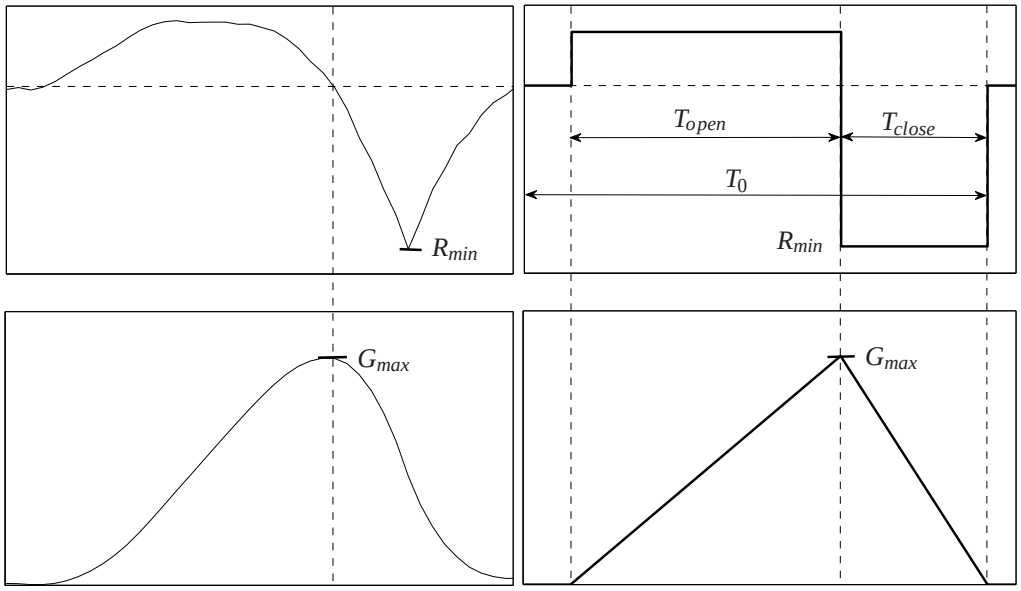
En la Figura 3.11(b) se presenta una estilización triangular del pulso glotal y su derivada (residuo) correspondiente. En esta estilización T_{close} es el tiempo que tarda el pulso en pasar de una amplitud G_{max} a 0. Puesto que se trata de una estilización lineal, este paso se realiza a una velocidad constante igual a R_{min}:

$$T_{close} = \frac{G_{max}}{R_{min}} \tag{3.29}$$

Sustituyendo en (3.28) tenemos la definición del NAQ:

$$NAQ = \frac{G_{max}}{T_0 \cdot R_{min}} \tag{3.30}$$

En realidad, el NAQ sólo es igual al CQ en el caso de la curva estilizada, no en la señal real. Sin embargo, se trata de una buena aproximación, estando los valores del NAQ y del CQ altamente correlados, tal y como demuestran Bäckström *et al.*



(A) *Señal glotal real* **(B)** *Señal estilizada*

FIGURA 3.11: *Valores asociados al cálculo del NAQ. Las figuras superiores muestran el resíduo LPC, mientras que las inferiores representan el pulso glotal correspondiente.*

(2002). Y lo que es más importante, su cálculo es mucho más robusto, ya que es más fácil medir las amplitudes del pulso glotal y del residuo en cada período que determinar con exactitud el instante de cierre de la glotis.

Pendiente espectral (tilt)

La *pendiente espectral* o *tilt* se define para las tramas sonoras de la señal, y se corresponde con la pendiente de la envolvente espectral del residuo LPC. Para calcular este valor se toma el residuo LPC de una trama, tal y como se obtiene del algoritmo de filtrado inverso, y se le aplica una DFT, obteniendo su espectro de potencia. Para estimar la pendiente de la envolvente se toma la potencia correspondiente a las frecuencias F_0 y sus armónicos y se calcula una regresión lineal sobre estos puntos. El tilt se define como la pendiente de esta recta de regresión, medida en dB por octava. Matemáticamente:

$$Tilt = \frac{\sum_{n=0}^{N}(f_n \cdot A_n) - \sum_{n=0}^{N} f_n \sum_{n=0}^{N} A_n}{\sum_{n=0}^{N} f_n^2 - \sum_{n=0}^{N} f_n \sum_{n=0}^{N} f_n} \tag{3.31}$$

siendo A_n la amplitud del armónico n-ésimo (en dB) y $f_n = \log_2((n+1)F_0)$ su valor frecuencial en octavas.

A modo de ejemplo la Figura 3.12 muestra un caso típico de espectro del residuo y el cálculo de la recta de regresión correspondiente. En la Figura 3.12(a) se muestra el espectro del residuo en dB/Hz, mostrando la caída exponencial de la envolvente. Convirtiendo el eje de frecuencias a octavas (Figura 3.12(b)) se convierte a una caída lineal, con lo que se puede representar fácilmente con una recta de regresión.

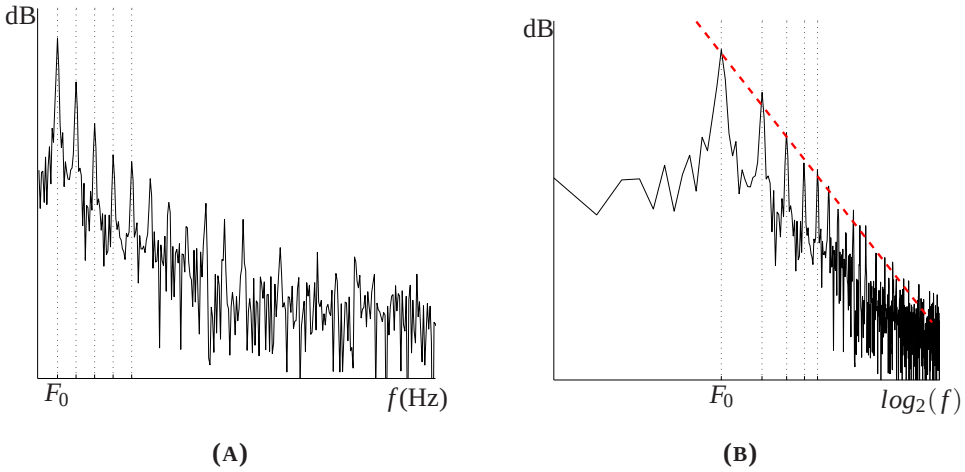


FIGURA 3.12: *Visualización gráfica de la pendiente espectral. (a) representa el espectro típico del residuo, con una envolvente exponencial decreciente. Convirtiendo el eje de frecuencias a octavas (b) la envolvente se transforma aproximadamente en una recta. Tomando los picos correspondientes a los armónicos de F_0 se puede calcular la envolvente como una recta de regresión (línea discontinua roja).*

Debido a las características del espectro del residuo, a altas frecuencias deja de haber componentes armónicas. Este efecto es más acusado si se tiene en cuenta que el residuo se ha calculado mediante un filtrado inverso no supervisado, con lo que su estimación contiene cierta componente de ruido. Además, cualquier error en la estimación del valor de F_0 se multiplica al calcular los armónicos altos, haciendo que los valores de amplitud asociados no se localicen correctamente. Por todo esto, para el cálculo del tilt sólo se han utilizado las amplitudes a frecuencia F_0 y sus primeros 4 armónicos ($N = 4$). Además, se han utilizado varias técnicas para incrementar la precisión en la detección de los armónicos y sus amplitudes.

La búsqueda de las amplitudes en los armónicos tiene dos fuentes principales de error: los errores en la estimación de F_0 y la precisión frecuencial de la DFT:

- Si la estimación de F_0 no es exacta, en los armónicos superiores este error se multiplica. Sea f_0 el valor correcto y $f_0 + \Delta$ el estimado. La posición estimada del cuarto armónico (el mayor utilizado en esta implementación) será $5f_0 + 5\Delta$. Si el valor 5Δ es superior al tamaño de un bin de la DFT, se estará leyendo la amplitud a una frecuencia incorrecta.

- Debido a la propia precisión espectral de la DFT, no habrá un bin centrado en el valor del armónico nf_0, por lo que no se puede estimar la amplitud exacta en ese armónico.

Para corregir en la medida de lo posible estos efectos, se aplica una sencilla técnica de búsqueda de máximos corregida con un ajuste parabólico:

- A partir del valor estimado para F_0 y sus armónicos se localiza el bin de la DFT correspondiente. A partir de este bin se busca hacia delante y atrás un máximo local, corrigiendo en parte la imprecisión del cálculo de F_0.

- Una vez localizado el máximo local, se toman las amplitudes de ese bin y de de los dos adyacentes y se aplica un ajuste parabólico a los 3 valores (Figura 3.13). De esta forma se puede localizar la posición del máximo y su valor de forma más precisa, corrigiendo los efectos de la precisión frecuencial de la DFT.

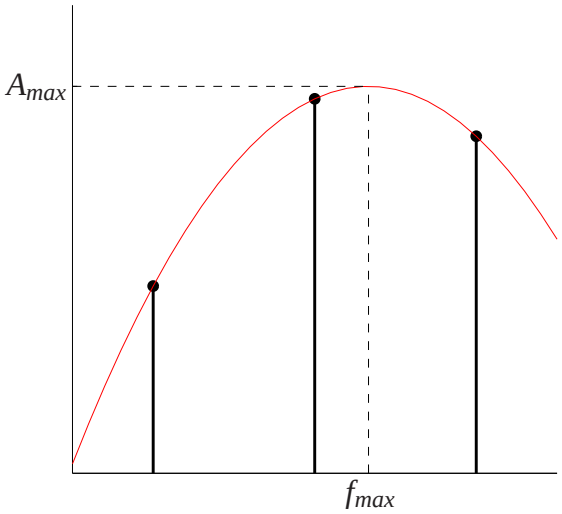


FIGURA 3.13: *Diagrama del ajuste parabólico aplicado a la búsqueda de máximos en el espectro glotal. Una vez localizado el bin de la DFT con la máxima potencia local, se toman los valores del bin anterior y posterior y se refina el valor del máximo y su posición gracias a la interpolación parabólica.*

Equilibrio espectral

El equilibrio espectral (SB, *Spectral Balance*) es una medida de la distribución de la potencia en el espectro de la señal. El valor de SB se puede calcular para cada trama a partir de la DFT (van Son y Pols, 1999):

$$SB = \frac{\sum_{i=1}^{k} f_i \cdot E_i}{\sum_{i=1}^{k} E_i} \tag{3.32}$$

donde E_i es la energía del bin i-ésimo de la DFT y f_i es la frecuencia central de ese bin. El SB representa por tanto el *centro de gravedad* del espectro, el punto que deja tanta potencia a frecuencias inferiores como superiores. Un valor de SB alto implica que hay más potencia a frecuencias altas mientras que un valor bajo representa mayor potencia a frecuencias bajas.

La relación entre el SB y la emoción de la voz ha sido comprobada por varios estudios (Kienast y Sendlmeier, 2000; Yildirim *et al.*, 2004), que revelan valores de SB mayores para las emociones de gran tensión (miedo, enfado) mientras que las de baja tensión (tristeza, aburrimiento) muestran un valor inferior.

3.3. Conclusiones

Se ha procurado que los parámetros considerados sean un reflejo de los más utilizados en la literatura para la identificación de emociones en el habla. Por ello se han seleccionado parametrizaciones de envolvente espectral, tanto a nivel de segmento como supra-segmentales, parámetros derivados de las curvas prosódicas y características de la calidad de voz. Además de las aquí definidas, podrían haberse considerado muchas otras características que ya han sido utilizadas en la investigación en el campo del habla emocional: características lingüísticas, formantes, correlaciones... Sin embargo, el número de alternativas es innumerable, y la capacidad de procesamiento limitada. Por ello se ha preferido limitar el estudio a las parametrizaciones que se consideran más representativas. El análisis de los parámetros definidos puede dar una idea bastante acertada de cómo se comportarán otros parámetros de naturaleza similar.

Capítulo 4

Análisis de los parámetros con emociones actuadas

Índice

ANALIZAR los parámetros en cuanto a su capacidad para discriminar emociones permite comprender el comportamiento de estas parametrizaciones cuando se utilizan para la identificación automática de emociones. La mayoría de los trabajos publicados en este campo analizan los parámetros de forma individual, dando lugar a conclusiones poco válidas cuando se consideran en conjunto. Muchos artículos publicados proporcionan tasas de precisión obtenidas mediante evaluación experimental para una cierta parametrización, lo que permite evaluar el comportamiento de todo el conjunto de características. Pero estos resultados no son comparables entre sí, debido a las diferencias en la arquitectura de los experimentos (número y tipo de emociones, dependencia o independencia de locutor, calidad de las grabaciones, etc.).

En este capítulo realiza un análisis de la capacidad de discriminación proporcionada por las diferentes parametrizaciones definidas en el capítulo 3. Este análisis se lleva a cabo tanto para las parametrizaciones individuales como para sus combinaciones, determinando de esta forma si estas combinaciones son provechosas o no. Además, también se efectúan pruebas empíricas de identificación de emociones para validar los resultados. Estas pruebas han sido realizadas bajo las mismas condiciones, utilizando la misma base de datos y arquitectura, por lo que los resultados presentados son totalmente comparables entre sí.

4.1. Descripción de las bases de datos de trabajo: *Berlin*

La base de datos *Berlin* (Burkhardt *et al.*, 2005) contiene 535 grabaciones de diez locutores, cinco hombres y cinco mujeres, simulando siete estados emocionales diferentes: aburrimiento, asco, enfado, felicidad, miedo, tristeza y el estilo neutro. Se trata, por tanto, de emociones actuadas. El corpus está formado por 10 frases de contenido semántico neutro. Todas las grabaciones están disponibles muestreadas a 16 kHz y 16 bits por muestra.

Los autores de esta base de datos pusieron mucho empeño en conseguir una alta naturalidad en las emociones recogidas. Varios locutores se presentaron a un proceso de selección en el que tres expertos evaluaron la naturalidad y facilidad de reconocimiento de las emociones que expresaban. De esta forma seleccionaron a los 10 locutores que forman parte de la base de datos. Tuvieron cuidado además de seleccionar cinco hombres y cinco mujeres, para mantener el equilibrio entre sexos.

Con el objetivo de reforzar la naturalidad, durante el proceso de grabación se hizo escuchar a los locutores una caracterización de la emoción a interpretar (por ejemplo, alguien que acababa de ganar la lotería para el caso de la felicidad). También se les pidió que rememoraran algún suceso de su vida asociado a esa emoción. Por último, se les dio una serie de indicaciones sobre cómo mejorar la naturalidad (como por ejemplo, no gritar para expresar el enfado). Todos los locutores grabaron las 10 frases en los 7 estilos considerados. Si en algún momento un locutor consideraba que la caracterización de la emoción no había sido correcta, se le permitía grabar la frase de nuevo.

Finalizado el proceso de grabación se obtuvieron cerca de 800 frases grabadas (incluyendo las repeticiones). Estas frases fueron objeto de una evaluación perceptual, donde 20 oyentes tenían que identificar la emoción expresada y puntuar su naturalidad. Se eliminaron todas las grabaciones que con una tasa de identificación de la emoción inferior al 80 % y una puntuación de naturalidad inferior al 60 %, quedando 535 frases en la base de datos definitiva. Por último, las grabaciones fueron etiquetadas a nivel de fonema y de sílaba de forma manual por expertos fonetistas.

A causa de la selección de las grabaciones mediante la evaluación perceptual, la base de datos no mantiene un equilibrio entre las emociones. En la Tabla 4.1 se muestra la distribución de las grabaciones, y puede comprobarse que el estilo enfadado está sobrerrepresentado, mientras que hay pocos ejemplos de asco.

Esta base de datos fue seleccionada porque presenta ciertas características interesantes. Por un lado se trata de una base de datos multilocutor, lo que permite realizar pruebas de identificación de emociones con una arquitectura independiente de locutor. Además, la selección de las frases mediante la evaluación perceptual garantiza cierta naturalidad en las emociones reflejadas, aun tratándose de actuaciones. Por último, se trata de una base de datos que ha sido utilizada en numerosos trabajos publicados (Chichosz y Slot, 2007; Lugger y Yang, 2007; Truong y van Leeuwen, 2007; Vogt y André, 2006), lo que permite obtener resultados comparativos con estos trabajos.

TABLA 4.1: *Distribución de las grabaciones en la base de datos* Berlin*.* Ab*: aburrimiento,* As*: asco,* En*: enfado,* Fe*: felicidad,* Mi*: miedo,* Ne*: neutro,* Tr*: tristeza.*

	Loc.	Ab	As	En	Fe	Mi	Ne	Tr	*Tot.*	*%*
Masc.	**L1**	5	1	14	7	4	11	7	*49*	*9,2*
	L2	8	1	10	4	8	4	3	*38*	*7,1*
	L3	8	2	11	8	10	9	7	*55*	*10,3*
	L4	5	2	12	2	6	4	4	*35*	*6,6*
	L5	9	5	13	6	8	11	4	*56*	*10,5*
Fem.	**L6**	10	0	12	11	6	10	9	*58*	*10,8*
	L7	4	8	13	4	1	9	4	*43*	*8,0*
	L8	10	8	12	10	7	9	5	*61*	*11,4*
	L9	8	8	16	8	12	7	10	*69*	*12,9*
	L10	14	11	14	11	7	5	9	*71*	*13,3*
	Tot.	*81*	*46*	*127*	*71*	*69*	*79*	*62*	*535*	*100*
	%	*15,1*	*8,6*	*23,7*	*13,3*	*12,9*	*14,8*	*11,6*	*100*	

4.2. Variabilidad inter-emoción e intra-emoción

La capacidad de una parametrización para retener las características de la emoción y evitar el resto de características asociadas a la voz puede medirse en términos de *dispersión intra-clase* y *dispersión inter-clase* (que en este caso se corresponden con *dispersión intra-emoción* y *dispersión inter-emoción* respectivamente):

- La **dispersión intra-clase** se define como la dispersión calculada sobre todas las muestras pertenecientes a una clase, lo que proporciona una medida del volumen que abarca la clase en el espacio vectorial de los parámetros. Cuando existe más de una clase, suele calcularse la dispersión media entre todas las clases para obtener un valor global.
- La **dispersión inter-clase** mide la dispersión entre los centroides de las clases, dando una estimación de lo separadas que están las clases unas de otras.

Sean S_W y S_B las matrices de dispersión intra-clase e inter-clase respectivamente:

$$S_W = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(y_i^j - \bar{y_i})(y_i^j - \bar{y_i})^T \tag{4.1}$$

$$S_B = \frac{1}{N}\sum_{i=1}^{M} N_i(\bar{y_i}-\bar{y})(\bar{y_i}-\bar{y})^T \tag{4.2}$$

donde N_i es el número de muestras disponibles para la clase i, M es el número de clases, $N = \sum_{i=1}^{M} N_i$ es el número total de muestras de entrenamiento e y_i^j son las muestras de la clase i. $\bar{y_i}$ es la media de la clase i mientras que $\bar{y}$ es la media global:

$$\bar{y_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_i^j \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_i^j \tag{4.3}$$

Una parametrización separará más fácilmente las emociones si tiene una dispersión intra-emoción pequeña y una dispersión inter-emoción grande. La relación entre estas dos dispersiones proporciona una medida del solapamiento de las distribuciones. En el caso multidimensional, esta medida puede estimarse utilizando el siguiente valor (Fukunaga, 1990):

$$J_1 = \mathrm{tr}(S_W^{-1}\cdot S_B) \tag{4.4}$$

donde $\mathrm{tr}(\cdot)$ representa la traza de una matriz. Esta medida se utiliza a menudo como criterio en análisis lineal discriminante (LDA, *Linear Discriminant Analysis*), y se trata de una generalización del conocido criterio de Fisher para el caso multiclase y multidimensional (Duda *et al.*, 2001):

$$J_F = \frac{|\bar{y_1}-\bar{y_2}|^2}{\sigma_1^2+\sigma_2^2} \tag{4.5}$$

El valor J_1 se ha utilizado como primera estimación de la capacidad de cada familia de parámetros para discriminar emociones. Los resultados obtenidos se detallan a continuación para los parámetros supra-segmentales y segmentales.

Parámetros supra-segmentales

Los resultados del cálculo de discriminalidad en las parametrizaciones supra-segmentales se muestran en la Tabla 4.2, tanto para cada familia de parámetros como para varias combinaciones de fusión temprana. La primera conclusión que se puede extraer de este cálculo es que los parámetros prosódicos presentan, en principio, una menor capacidad para separar las emociones que los estadísticos de la envolvente espectral, siendo su valor de J_1 aproximadamente cinco veces menor. Esto es especialmente significativo, ya que las características prosódicas son con diferencia las más utilizadas en la literatura (ver Tabla 2.2). Este resultado puede ser parcialmente justificado por el mayor número de parámetros utilizados

en el caso espectral (324 parámetros espectrales frente a 54 prosódicos). Sin embargo, cuando los cálculos se repiten utilizando sólo los mejores 54 parámetros espectrales proporcionados por el algoritmo de selección (ver sección 4.4), se obtiene un valor de discriminalidad de $9,12$, todavía superior al de los parámetros prosódicos.

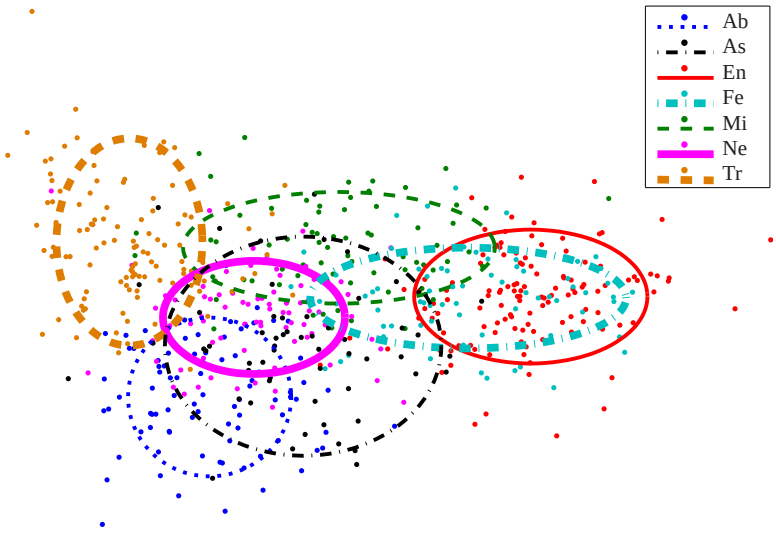
TABLA 4.2: *Discriminalidad de los parámetros supra-segmentales.*

Parámetros	J_1
Prosódicos	$6,47$
Calidad	$0,99$
Espectrales	$34,43$
Pros.+Calidad	$7,14$
Pros.+Calidad+Espec.	$54,36$

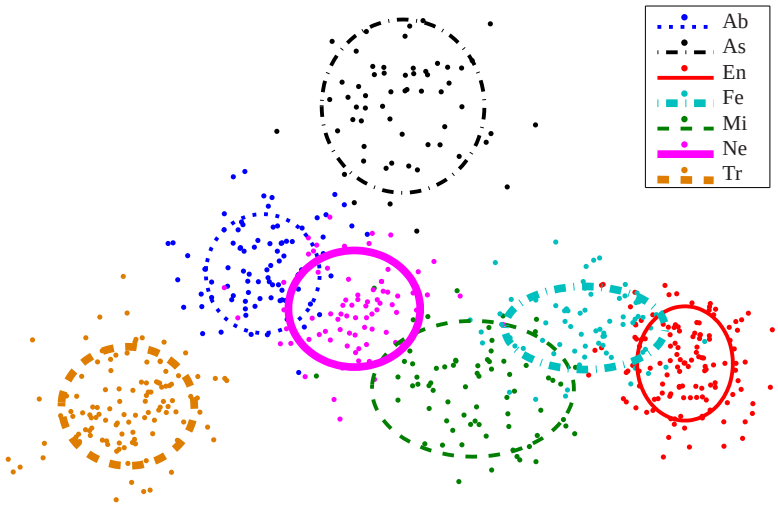
Con el objetivo de mostrar la capacidad de discriminación de los parámetros de forma visual y confirmar estos resultados se ha realizado un análisis LDA utilizando tanto características prosódicas como espectrales. A partir de este análisis se han localizado las dos direcciones más discriminantes en el espacio de parámetros y se han proyectado los vectores de características al plano definido por estas dos direcciones. La distribución resultante puede verse en la Figura 4.1. Como puede apreciarse, las emociones aparecen menos solapadas con la parametrización espectral (Figura 4.1(b)) que con la prosódica (Figura 4.1(a)). Curiosamente, en ambos casos la dirección más discriminante (eje horizontal) parece estar muy relacionada con el nivel de tensión de la emoción, desplazando las emociones de alta tensión (enfado, alegría y miedo) hacia un lado y las de poca tensión (tristeza y aburrimiento) hacia el otro.

En el diagrama relativo a las características prosódicas también pueden detectarse las confusiones más habituales descritas en la literatura al utilizar este tipo de parametrización. Un caso típico es la confusión entre el enfado y la alegría, ambas de alta tensión. El asco es una emoción tradicionalmente difícil de detectar mediante parámetros prosódicos (Burkhardt y Sendlmeier, 2000; Iriondo *et al.*, 2000), lo que también puede apreciarse en el diagrama. La distribución del asco tiene una elevada dispersión y está fuertemente solapada con varias emociones, sobre todo con el estilo neutro. Sin embargo, los parámetros espectrales muestran un comportamiento diferente, separando el asco correctamente del resto de emociones.

Volviendo a los valores de discriminalidad mostrados en la Tabla 4.2, se puede comprobar que los parámetros de calidad de voz parecen proporcionar muy poca



(A) *Dispersión de clases con parámetros prosódicos.*



(B) *Dispersión de clases con parámetros espectrales.*

FIGURA 4.1: *Diagrama de dispersión de los parámetros supra-segmentales prosódicos (a) y espectrales (b), proyectados sobre las dos direcciones más discriminantes según el análisis LDA. Las elipses delimitan la región a dos desviaciones estándar del centroide.* En*: enfado,* Mi*: miedo,* Ab*: aburrimiento,* Ne*: neutro,* Fe*: felicidad,* Tr*: tristeza,* As*: asco.*

información, al menos los utilizados en este análisis. Sin embargo, el uso de estos parámetros no es del todo inútil. Combinándolos con los parámetros prosódicos se consigue incrementar la discriminalidad de un 6,47 a un 7,14. Este fenómeno puede explicarse teniendo en cuenta que parámetros que por sí mismos no tienen capacidad de discriminación pueden ser útiles cuando se combinan con otros (Guyon y Elisseeff, 2003), como parece ser el caso. Por último, la concatenación de todos los parámetros supra-segmentales proporciona la máxima discriminación, lo que sugiere que, efectivamente, la información recogida por la envolvente espectral, la prosodia y la calidad de voz es complementaria.

Parámetros segmentales

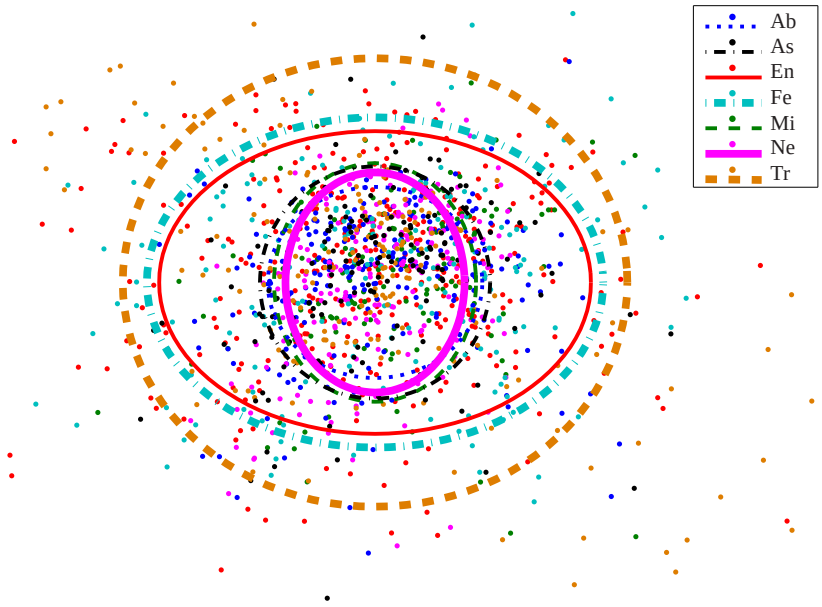
La Tabla 4.3 muestra los valores de J_1 para cada parametrización considerada a nivel de segmento. Según estos valores, la separación que proporcionan estas parametrizaciones entre las emociones es prácticamente nula. Este resultado también puede comprobarse en las Figuras 4.2(a) y 4.2(b), que muestran las dos direcciones más discriminantes según el análisis LDA para primitivas de prosodia y LFPC respectivamente. Como puede verse, las distribuciones están completamente solapadas.

TABLA 4.3: *Discriminalidad de los parámetros segmentales.*

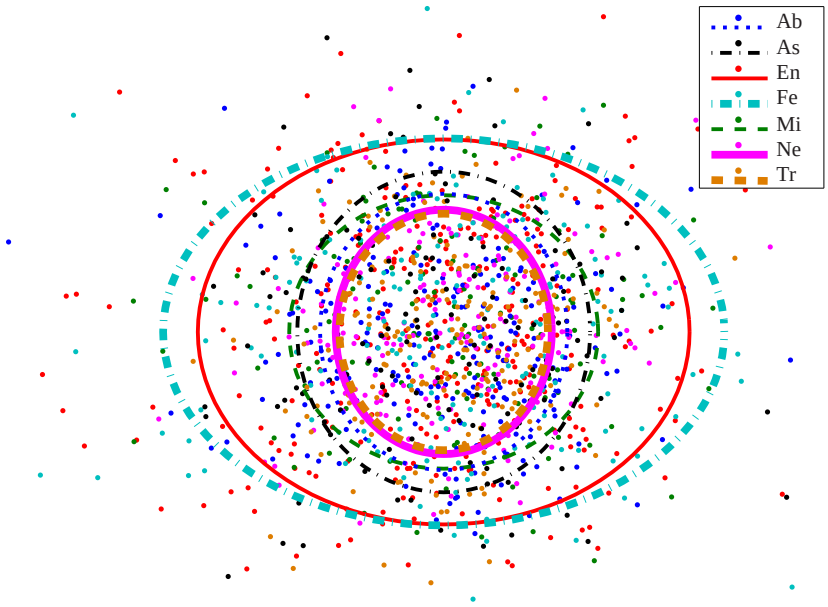
Parámetros	J_1
LFPC	$1,98 \cdot 10^{-10}$
Prim. pros. - sonoras	$1,83 \cdot 10^{-11}$
Prim. pros. - sordas	$8,75 \cdot 10^{-13}$

Este solapamiento ocurre debido a la propia naturaleza de los parámetros a corto plazo. Por ejemplo, cada vector LFPC refleja la envolvente espectral de una única trama, es decir, representa la envolvente espectral característica del fonema que se está articulando en esa trama. Puesto que la envolvente espectral difiere mucho más entre diferentes fonemas que entre diferentes emociones, la dispersión intra-clase es muy grande, provocando un gran solapamiento entre las emociones. En el caso de las primitivas de prosodia ocurre un efecto similar, ya que las muestras a nivel de trama recogen una mayor variación debido al contenido lingüístico que al contenido emocional de la frase.

Esto no significa que las parametrizaciones a corto plazo no puedan ser utilizadas para la identificación de emociones. Tanto el criterio J_1 como el análisis LDA son óptimos sólo si las clases tienen distribuciones normales homoscedásticas,



(A) *Dispersión de las clases con primitivas de prosodia.*



(B) *Dispersión de las clases con parámetros LFPC.*

FIGURA 4.2: *Diagrama de dispersión de las primitivas de prosodia (a) y LFPC (b), proyectados sobre las dos direcciones más discriminantes según el análisis LDA. Las elipses delimitan la región a dos desviaciones estándar del centroide.* En*: enfado,* Mi*: miedo,* Ab*: aburrimiento,* Ne*: neutro,* Fe*: felicidad,* Tr*: tristeza,* As*: asco.*

es decir, cuando tienen la misma matriz de covarianzas (Duda *et al.*, 2001). Por tanto, estos métodos no pueden detectar las pequeñas diferencias existentes en la forma de las distribuciones. Sin embargo, estas diferencias pueden ser capturadas utilizando sistemas de clasificación adecuados. Para ello se suelen utilizar GMM, que son capaces de aprovechar estas pequeñas diferencias en las distribuciones, siempre y cuando se proporcionen suficientes muestras de entrenamiento. Puesto que en las parametrizaciones segmentales se extrae un nuevo vector de parámetros cada 10 ms, proporcionan un elevado número de muestras de entrenamiento por cada frase, por lo que es posible entrenar GMM precisos y robustos. De hecho, los GMM son muy utilizados en sistemas de identificación de emociones en los que se utilizan parámetros a nivel de trama. Desgraciadamente, el gran solapamiento existente entre las clases significa que no es posible extraer conclusiones razonables para las parametrizaciones segmentales a partir de las medidas de J_1. La única forma de obtener una estimación de su capacidad para distinguir emociones es mediante pruebas empíricas de identificación automática. Estas pruebas empíricas se presentan en la sección 4.5.

4.3. Agrupación no supervisada

Los valores J_1 proporcionan una idea de la capacidad de discriminación de cada familia de parámetros, mientras que el diagrama de dispersión de las direcciones más discriminantes facilita una representación visual del mismo dato. Esta representación visual es incompleta y poco realista, ya que incluyendo más dimensiones se puede alcanzar una mayor separación entre las clases. Por desgracia, no es posible realizar una representación gráfica de todas estas dimensiones. Sin embargo, se pueden obtener resultados descriptivos que proporcionen cierta información de lo que ocurre cuando se utilizan todos los parámetros, sin tener que llegar a diseñar un sistema completo de identificación automática.

Para llevar a cabo este análisis se ha realizado una agrupación ciega no supervisada de los vectores de parámetros utilizando el algoritmo k-means (Duda *et al.*, 2001). Si las clases se encuentran correctamente separadas, los conjuntos obtenidos se deberían corresponder con cada emoción. El resultado de esta agrupación se presenta en las Tablas 4.4 y 4.5 para el caso de los parámetros supra-segmentales. No se realizó ninguna agrupación para los parámetros segmentales, ya que las distribuciones están tan solapadas que el algoritmo no podría hallar las clases correctamente.

Se puede comprobar que el algoritmo de agrupación es capaz de identificar las muestras pertenecientes a cada emoción con bastante precisión. Coincidiendo con los resultados de las medidas de J_1, la parametrización espectral permite identificar las clases mejor que la parametrización prosódica. Con los estadísticos de

TABLA 4.4: *Resultados del clustering ciego para parámetros prosódicos.*

	C1	C2	C3	C4	C5	C6	C7
Aburrimiento	71	1	0	0	1	9	2
Asco	5	40	1	1	1	8	3
Enfado	1	0	100	23	3	1	0
Felicidad	0	2	12	51	3	4	0
Miedo	0	2	3	4	46	15	2
Neutro	9	1	1	1	0	66	1
Tristeza	4	0	0	0	3	19	85

TABLA 4.5: *Resultados del clustering ciego para estadísticos de espectro.*

	C1	C2	C3	C4	C5	C6	C7
Aburrimiento	81	0	81	0	0	3	0
Asco	0	59	0	0	0	0	0
Enfado	0	0	128	0	0	0	0
Felicidad	0	0	3	69	0	0	0
Miedo	0	0	0	0	72	0	0
Neutro	0	0	0	0	1	78	0
Tristeza	1	0	0	0	0	0	111

espectro casi todas las muestras de una cierta clase han sido asignadas al mismo conjunto, mientras que los parámetros prosódicos muestran las confusiones típicas entre emociones: enfado con felicidad y neutro con aburrimiento y tristeza. Si se considerara estas tablas como matrices de confusión, la precisión de la parametrización prosódica sería del 75,87 % mientras que la de los estadísticos de espectro llegaría al 98,68 %. Hay que notar que la precisión esperada para un sistema de identificación de emociones sería en realidad mucho menor, ya que las señales de prueba no deberían haber sido vistas durante el entrenamiento.

4.4. Selección de parámetros

El uso de parámetros redundantes que no aportan información puede provocar una reducción en la precisión del clasificador, debido a la confusión que añaden al sistema. Por ello, en los sistemas de clasificación es muy habitual utilizar algún mecanismo de selección de parámetros que permita separar aquellas característi-

cas que realmente aportan información discriminante de las que sólo añaden ruido a la parametrización. Con ello también se consigue reducir la dimensionalidad de la parametrización, lo que hace que el proceso de clasificación tenga menor carga computacional. El detectar los parámetros verdaderamente discriminantes puede permitir además comprender mejor la relación entre las emociones y las características acústicas de la voz.

Los algoritmos de selección de parámetros se suelen dividir en dos grupos, los *wrappers* y los *filtros* (Guyon y Elisseeff, 2003). Los métodos wrapper utilizan un clasificador como mecanismo para estimar la precisión que puede alcanzar un cierto conjunto de parámetros. Esta estimación se obtiene entrenando y evaluando el clasificador con cada conjunto de parámetros considerado, generalmente mediante algún tipo de validación cruzada. Una vez estimado el error de clasificación cometido por cada conjunto de parámetros, se selecciona aquél que proporciona mayor tasa de aciertos. Este tipo de métodos suele proporcionar conjuntos de parámetros que alcanzan una gran precisión para el clasificador considerado. Sin embargo, los resultados no suelen ser completamente generalizables a otros clasificadores. Además, son algoritmos de gran carga computacional, debido a la necesidad de entrenar el clasificador por cada conjunto de parámetros analizado.

En cambio, los métodos filtro seleccionan los parámetros en función de su distribución y su relación con el resto de parámetros y con la clase objetivo. Como resultado, el conjunto seleccionado no está optimizado para ningún clasificador, y generalmente proporciona una precisión comparable con diferentes clasificadores. También requieren una carga computacional significativamente menor, al no tener que entrenar un clasificador en cada etapa.

En este caso se ha utilizado el algoritmo mínima-redundancia-máxima-relevancia (mRMR, *Minimal-Redundancy-Maximal-Relevance*) (Peng *et al.*, 2005) para ordenar los parámetros desde el más significativo al menos significativo. Se trata de un algoritmo de tipo filtro, que selecciona aquellos parámetros que maximizan la información mutua entre las muestras de entrenamiento y su clase correspondiente (máxima relevancia), a la vez que minimiza la dependencia mutua entre los parámetros ya seleccionados (mínima redundancia). Se ha aplicado este método a las cinco parametrizaciones definidas en la sección 3.2 (supra-segmentales: estadísticos de envolvente espectral, prosodia y calidad de voz; segmentales: LFPC y primitivas de prosodia), así como a sus diferentes combinaciones de fusión temprana.

La Tabla 4.6 muestra los 30 primeros parámetros seleccionados en el caso de las combinaciones. Estas listas son muy significativas, ya que permiten comprobar qué parametrizaciones son preferidas en cada caso. Por ejemplo, cuando se concatenan todos los parámetros supra-segmentales, entre los 20 primeros encontramos seis parámetros prosódicos y 14 espectrales. Si aumentamos un poco el rango, entre los 30 primeros parámetros encontramos diez prosódicos, uno de ca-

TABLA 4.6: *Los primeros 30 parámetros seleccionados en las parametrizaciones combinadas.* VQ*: calidad de voz,* PP*: primitivas de prosodia.*

#	Parámetros supra-segmentales		Parámetros segmentales	
	Prosodia+ VQ	**Prosodia+ VQ+ Espectral**	**LFPC+ PP sonoros**	**LFPC+ PP sordos**
1	R(Pow)	min($LFPC_1$)	ΔF_0	$\Delta LFPC_{14}$
2	K($\Delta^2 F_0$)	Sk($LFPC_2$)	$\Delta^2 LFPC_3$	Pow
3	K(ΔPow)	E($\Delta^2 LFPC_7$)	F_0	$\Delta^2 LFPC_{15}$
4	$LvVdur$	E($LFPC_{13}$)	$\Delta LFPC_1$	$\Delta LFPC_{18}$
5	R(ΔF_0)	K($\Delta LFPC_7$)	$LFPC_{15}$	$\Delta^2 LFPC_8$
6	$NLvPcn$	$NLvPcn$	$\Delta^2 F_0$	ΔPow
7	E(Pow)	E($\Delta LFPC_1$)	$LFPC_2$	$LFPC_{14}$
8	Shm	E($Vdur$)	$\Delta^2 LFPC_1$	$\Delta LFPC_{17}$
9	$\sigma^2(\Delta^2 Pow)$	E(Pow)	$\Delta LFPC_{18}$	$\Delta^2 LFPC_{13}$
10	K(Pow)	max($F_0 sl$)	$\Delta LFPC_2$	$LFPC_1$
11	min(F_0)	$\sigma^2(LFPC_{17})$	$LFPC_3$	$\Delta LFPC_7$
12	E($Vdur$)	Sk($\Delta^2 LFPC_1$)	$\Delta^2 LFPC_8$	$\Delta^2 LFPC_5$
13	SB	Sk($LFPC_{16}$)	ΔPow	$\Delta LFPC_{10}$
14	$\sigma^2(Psl)$	K($\Delta LFPC_{18}$)	$LFPC_4$	$\Delta^2 LFPC_{16}$
15	$LvF_0 sl$	Sk($\Delta LFPC_{12}$)	$LFPC_1$	$\Delta LFPC_{15}$
16	Sk(ΔPow)	E($LFPC_1$)	$\Delta^2 LFPC_{14}$	$\Delta^2 LFPC_9$
17	R(F_0)	R($\Delta^2 LFPC_{12}$)	$\Delta^2 LFPC_2$	$\Delta LFPC_4$
18	$LvPsl$	K($LFPC_8$)	$LFPC_9$	$\Delta LFPC_9$
19	E(ΔPow)	K($\Delta^2 Pow$)	$\Delta LFPC_3$	$\Delta^2 LFPC_{14}$
20	min(Pow)	E(Psl)	$\Delta LFPC_{14}$	$\Delta LFPC_{16}$
21	K($\Delta^2 Pow$)	Sk(F_0)	Pow	$LFPC_9$
22	max($F_0 sl$)	min($\Delta LFPC_6$)	$\Delta^2 LFPC_{16}$	$\Delta^2 Pow$
23	TLT	SB	$\Delta^2 Pow$	$\Delta^2 LFPC_{12}$
24	$NlvF_0 cn$	$NlvVdur$	$\Delta LFPC_5$	$\Delta LFPC_{12}$
25	$\sigma^2(Pow)$	R($LFPC_3$)	$\Delta LFPC_{18}$	$\Delta^2 LFPC_7$
26	Jit	E($\Delta LFPC_{17}$)	$\Delta LFPC_4$	$\Delta LFPC_{18}$
27	E($\Delta^2 Pow$)	min(Pow)	$\Delta^2 LFPC_{18}$	$\Delta^2 LFPC_{17}$
28	E(F_0)	K($\Delta LFPC_1$)	$LFPC_{14}$	$\Delta LFPC_6$
29	$\sigma^2(\Delta Pow)$	R(F_0)	$\Delta LFPC_9$	$LFPC_{18}$
30	Sk(F_0)	E($LFPC_{18}$)	$\Delta^2 LFPC_9$	$\Delta LFPC_{10}$

lidad de voz y 19 espectrales. Estos resultados son similares a los obtenidos por Schuller *et al.* (2005) y Vogt y André (2006), y refuerzan las conclusiones de la sección 4.2, que sugerían que los estadísticos de espectro proporcionan más información que los prosódicos acerca de la emoción. Los parámetros de calidad de voz quedan muy por debajo en la lista, aunque varios estudios descritos en la literatura manifiestan su relación con el estado emocional del locutor (Gobl y Chasaide, 2003; Lugger y Yang, 2007). Probablemente la mala posición de las características de calidad de voz se deba a los métodos utilizados para su extracción. En los trabajos que utilizan la calidad de voz, la parametrización se realiza generalmente con intervención humana, lo que permite *corregir* los errores de cálculo. Sin embargo, en el estudio presentado en este documento todo el proceso se ha realizado de forma automática. Aunque las características de calidad de voz se han extraído sólo en las vocales (que se supone que son segmentos muy estables de la señal), los errores en la estimación de estas características pueden haber incrementado la confusión entre las emociones. A estos errores hay que añadir además los que posiblemente se hayan cometido en la identificación automática de las vocales. Todo esto puede hacer que los parámetros de calidad de voz utilizados no sean adecuados para la clasificación de las emociones, tal y como se observa no sólo en las bajas posiciones del ranking, sino también en el valor de J_1 obtenido por esta parametrización.

La selección de parámetros se ha realizado de forma independiente para los flujos sordo y sonoro de las parametrizaciones segmentales. Aplicando un proceso similar al utilizado para las primitivas de prosodia, los LFPC también se han separado en dos flujos, uno para las tramas sonoras y otro para las sordas, de forma que la parametrización espectral a corto plazo y las primitivas de prosodia puedan combinarse fácilmente mediante concatenación. En la combinación de los flujos de tramas sonoras, los tres parámetros de entonación (F_0, ΔF_0 y $\Delta^2 F_0$) están entre los 10 mejores, mientras que los de intensidad (*Pow*, Δ*Pow* y Δ^2*Pow*) están entre las posiciones 10 y 25. En el flujo de tramas sordas estos valores de intensidad aparecen en los primeros puestos. El hecho de que los valores derivados de la curva de entonación se sitúen en las primeras posiciones corrobora que estos parámetros proporcionan mucha información acerca de la emoción presente en la voz.

4.5. Evaluación experimental

4.5.1. Marco experimental

Para validar los resultados obtenidos en el análisis de los parámetros, se han realizado pruebas de identificación automática de emociones sobre la propia base

de datos *Berlin*. Puesto que esta base de datos no está equilibrada en cuanto al número de señales disponible para cada emoción, se ha utilizado la precisión media no ponderada (UAR, *Unweighted Average Recall*) como medida de precisión, en lugar de la más tradicional precisión media ponderada (WAR, *Weighted Average Recall*)[1]. Mientras que la UAR se calcula promediando la precisión obtenida para cada emoción, la WAR se define como la tasa de acierto global de todas las pruebas realizadas.

$$UAR = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{N}\sum_{i=1}^{N} \frac{A_i}{M_i} \tag{4.6}$$

$$WAR = \frac{\sum_{i=1}^{N} A_i}{\sum_{j=1}^{N} M_j} = \sum_{i=1}^{N} \frac{A_i \cdot M_i}{M_i \cdot \sum_{j=1}^{N} M_j} = \sum_{i=1}^{N} P_i \cdot \Pr\{c = i\} \tag{4.7}$$

siendo N el número de emociones, P_i la precisión para la emoción i, M_i el número de señales de prueba para esta emoción y A_i el número de señales de prueba de la emoción i correctamente clasificadas. $\Pr\{c = i\} = \frac{M_i}{\sum_{j=1}^{N} M_j}$ representa la probabilidad *a priori* de que una señal de prueba pertenezca a la clase i. Por tanto, la WAR es equivalente a la media ponderada de las precisiones para cada emoción, utilizando la probabilidad *a priori* de cada emoción como peso de ponderación. La medida proporcionada por la UAR es más significativa en el caso de tener muestras desequilibradas, ya que considera todas las emociones equiprobables a la hora de realizar este cálculo. Por ejemplo, en un caso con dos emociones c_1 y c_2, con $\Pr\{c_1\} = 0,9$ y $\Pr\{c_2\} = 0,1$, un clasificador trivial que siempre devuelva la etiqueta c_1 obtendría una WAR del 90 %, mientras que la UAR sería del 50 %.

Clasificadores utilizados

Los parámetros supra-segmentales se han modelado mediante SVM con kernel RBF, mientras que para los parámetros segmentales se han utilizado GMM. De esta forma se pueden aprovechar las características de cada parametrización. Por un lado, los GMM pueden utilizar el elevado número de muestras de entrenamiento proporcionado por las parametrizaciones segmentales para modelar las diferencias en las distribuciones, las cuales están fuertemente solapadas. Por otro, las SVM pueden aprovechar la mayor separación entre distribuciones de parámetros supra-segmentales. Además, la alta capacidad de generalización proporcionada por las SVM (Burges, 1998) permiten entrenar modelos robustos incluso con el reducido número de vectores de entrenamiento proporcionados por estas parametrizaciones.

[1] Generalmente, cuando se dan valores de precisión sin especificar si son UAR o WAR, el dato proporcionado se suele referir a esta última.

Puesto que la evaluación de la precisión se realiza en términos de UAR, se ha implementado un clasificador GMM que considera todas las emociones equiprobables a la hora de tomar la decisión. Esto es equivalente a no utilizar la probabilidad *a priori* en la expresión (2.5):

$$\hat{c} = \arg\max_c P(x|c)\frac{1}{N} = \arg\max_c P(x|c) \tag{4.8}$$

siendo N el número de emociones consideradas. Gracias a esta implementación, no se favorece la clasificación en emociones sobrerrepresentadas.

Para las pruebas con modelos SVM se ha utilizado la biblioteca de funciones libSVM (Chang y Lin, 2004). El problema de la clasificación multiclase se ha resuelto mediante la aproximación uno contra todos (Hsu y Lin, 2002). Además, teniendo en cuenta que no todas las emociones están igualmente representadas, se ha utilizado la técnica de ponderación de costes (Osuna *et al.*, 1997) para compensar la falta de muestras de entrenamiento de algunas emociones. Esta técnica permite especificar diferentes costes de clasificación errónea a cada clase. Cuando el coste se calcula inversamente proporcional al número de muestras de entrenamiento, las clases menos representadas son favorecidas, compensando la falta de representación.

División de la base de datos

Durante las pruebas se ha tenido especial cuidado de que los resultados obtenidos sean independientes de locutor. Para ello, los experimentos se han diseñado como una doble validación cruzada anidada (ver Figura 4.3). El primer nivel divide la base de datos en cinco bloques, definidos en función de los locutores, de forma que proporciona pruebas independientes de locutor. El segundo nivel se utiliza para obtener resultados de desarrollo, necesarios para la optimización de los clasificadores.

Para el primer nivel de validación cruzada, los diez locutores de la base de datos se han dividido en cinco bloques $X = \{x_1, \ldots, x_5\}$. Cada bloque contiene un hombre y una mujer, manteniendo así el equilibrio de sexos entre los mismos (Figura 4.3(a)). En la etapa i-ésima del primer nivel, los bloques $TRAIN_i = X - x_i$ definen el conjunto de entrenamiento, dejando el bloque x_i para las pruebas. Para generar resultados de desarrollo para esta etapa, las señales de $TRAIN_i$ se distribuyen aleatoriamente en cinco sub-bloques $Y = \{y_{i1}, \ldots, y_{i5}\}$, que son utilizados para el segundo nivel de validación cruzada. Estos resultados de desarrollo se utilizan para estimar los valores óptimos del número de mezclas de los GMM, la dispersión del kernel RBF, el coste de clasificación errónea de la SVM y el número óptimo de parámetros. Una vez determinados estos valores, todas las señales

de $TRAIN_i$ se utilizan para el entrenamiento de la etapa i-ésima, aplicando finalmente el modelo sobre las señales de prueba del bloque x_i.

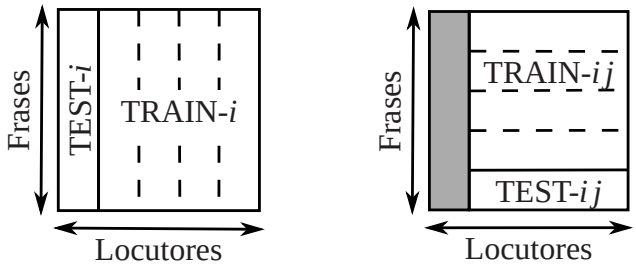


FIGURA 4.3: *Doble validación cruzada anidada aplicada sobre la base de datos* Berlin. *Para el primer nivel (a) se definen 5 bloques de dos locutores cada uno. Para el segundo nivel (b), las señales de entrenamiento se dividen en cinco subbloques de forma aleatoria.*

4.5.2. Selección del número de parámetros

Para estimar el número óptimo de parámetros a utilizar en cada parametrización, las pruebas de desarrollo se han repetido añadiendo un parámetro más cada vez, según el ranking obtenido en la sección 4.4. La Figura 4.4 muestra la precisión de las pruebas de desarrollo con parámetros supra-segmentales en función del número de parámetros.

Según estos resultados, aunque las características de calidad de voz proporcionan una reducida separación entre clases (ver Tabla 4.2), no se comportan tan mal considerando el reducido número de parámetros. Cuando se utilizan las cinco características de calidad de voz definidas, el sistema logra un 49,4 % de aciertos. Si se utilizan los cinco mejores parámetros prosódicos se consigue un resultado muy similar (50,7 %). Además, la combinación de parámetros prosódicos y de calidad de voz parece ser útil, tal y como predecían las medidas de separación de clases: el mejor sistema prosódico alcanza un máximo de precisión de 65,5 % con 39 parámetros, mientras que la combinación llega al 67,4 % con 17 parámetros. No sólo se incrementa la precisión, sino que además se alcanza el máximo con un número de parámetros menor.

Los estadísticos espectrales a largo plazo superan claramente a los prosódicos, al menos si se utilizan más de 15 parámetros. Esto confirma otra vez las conclusiones obtenidas mediante el análisis de separación de clases. La precisión para los estadísticos de espectro se satura alrededor de 96 parámetros, con un 75,4 %. Por último, la fusión temprana de todas las parametrizaciones supra-segmentales

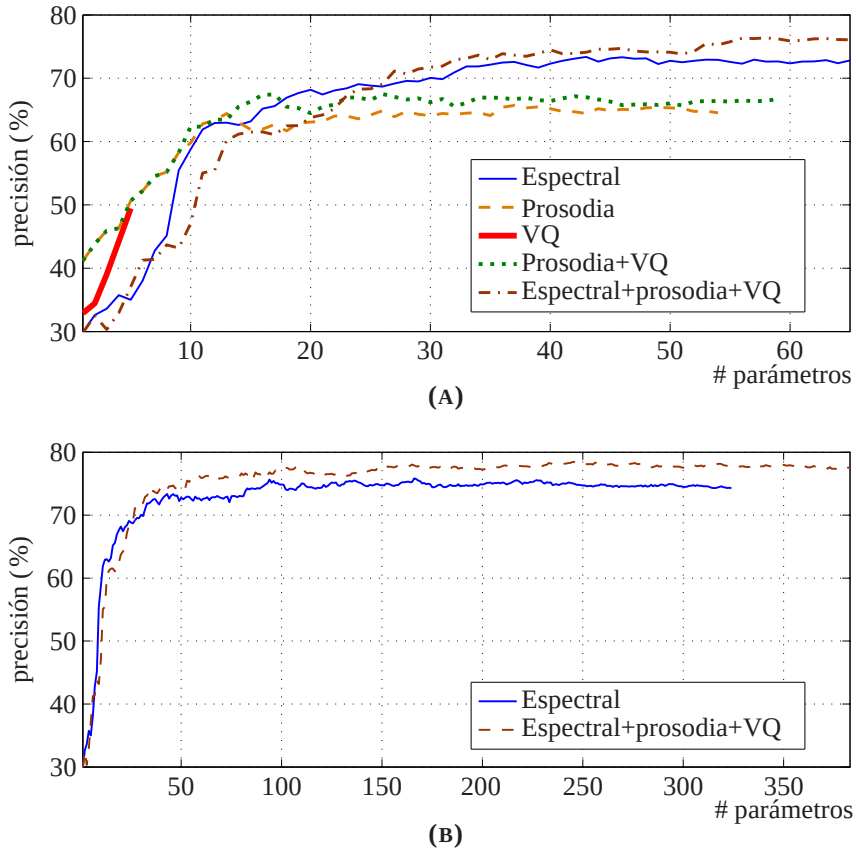


FIGURA 4.4: *Resultados de las pruebas de desarrollo para parametrizaciones supra-segmentales. Los gráficos muestran la precisión del sistema en función del número de parámetros. (a) Comparativa entre todas las parametrizaciones analizadas. (b) Reducción de escala para ver el comportamiento de las parametrizaciones con más de 60 parámetros.*

obtiene los mejores resultados para más de 25 parámetros, alcanzando un valor aproximadamente estable con 152 parámetros (77,9 %) y un máximo absoluto marginalmente superior con 247 (78,6 %).

Los resultados de desarrollo para los parámetros segmentales se presentan en la Figura 4.5. Ninguna de las curvas alcanza un verdadero punto de saturación, ya que la precisión sigue creciendo según se van añadiendo nuevos parámetros. Sin embargo, puede observarse que la mejora es pequeña a partir de 20 parámetros. Por ejemplo, con la parametrización LFPC se alcanza un 70,9 % de precisión con 20 parámetros, mientras que cuando se utilizan todos los 54 esta precisión se incrementa sólo hasta un 72,9 %. Cuando los flujos sordos y sonoros se consideran por separado, la precisión de los LFPC se reduce, lo cual es razonable, ya que en este caso hay aproximadamente la mitad de muestras de entrenamiento para cada

flujo.

Los resultados con las primitivas de prosodia son bastante modestos, alcanzando sólo un 65,3 % de precisión en el flujo sonoro y un 50,8 % en el sordo. Sin embargo, es necesario tener en cuenta que tan sólo utilizan seis y tres parámetros respectivamente. Por ejemplo, si se utilizan tan sólo los mejores seis parámetros del flujo sonoro de LFPC, el sistema logra sólo un 58,8 % de aciertos. Por tanto, para un número tan reducido de parámetros, las primitivas de prosodia tienen un buen comportamiento. Por el contrario, añadir estas primitivas a los vectores de LFPC no mejora los resultados significativamente. Según el ranking mostrado en la Tabla 4.6, las primitivas de prosodia se seleccionan entre los primeros 10 ó 20 parámetros, sugiriendo que son más informativas que la mayoría de los LFPC. Sin embargo, ésto sólo es cierto cuando se utilizan pocos parámetros (en estas pruebas, menos de 15). Cuando el número de parámetros utilizados aumenta, los nuevos LFPC añadidos compensan la información proporcionada por las primitivas de prosodia, de forma que al final la combinación no tiene ningún efecto sobre la precisión global.

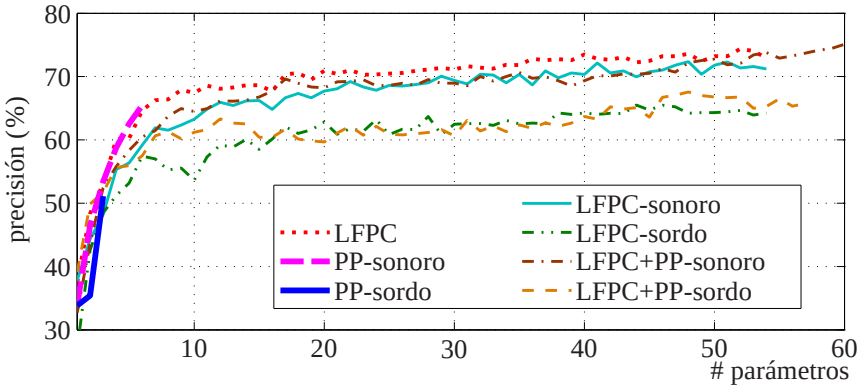


FIGURA 4.5: *Resultados de las pruebas de desarrollo para las parametrizaciones segmentales. El gráfico muestra la precisión del sistema en función del número de parámetros.*

La Tabla 4.7 resume los resultados de desarrollo para cada parametrización y sus combinaciones por fusión temprana. Los resultados se dan tanto para el número de parámetros óptimo estimado como para el conjunto completo de parámetros. También se muestran los resultados finales de las pruebas independientes de locutor para cada caso. En vista de estos resultados finales, se puede concluir que los estadísticos espectrales son la mejor parametrización aislada, alcanzando un 70,5 % de acierto con 96 parámetros (y prácticamente la misma precisión cuando se utilizan todos los 324 estadísticos definidos). Dentro de las combinaciones, el mejor resultado se obtiene con la combinación de todos los parámetros suprasegmentales, que proporciona un 72,2 % de aciertos con 152 parámetros.

4.5.3. Fusión tardía de expertos

Hasta el momento se han analizado los resultados de la fusión temprana, es decir, la simple concatenación de los vectores de parámetros. Pero también es interesante comprobar si los resultados varían al aplicar una fusión tardía, no combinando los parámetros directamente, sino los resultados de los clasificadores entrenados con cada parametrización. Además, la fusión tardía permite combinar la información capturada por parametrizaciones de diferente estructura temporal, tales como parámetros segmentales y supra-segmentales o los flujos de tramas sordas y sonoras.

Para llevar a cabo este proceso se ha utilizado un sistema de fusión basado en SVM (Fierrez-Aguilar *et al.*, 2005; Gutschoven y Verlinde, 2000). Sea $Q = \{q_1, \ldots, q_M\}$ un conjunto de M clasificadores y $C = \{c_1, \ldots, c_N\}$ el conjunto de N clases consideradas. Dada una señal a clasificar x, es posible recuperar una puntuación o *score* s_i^j que representa la confianza estimada por el clasificador q_j de que esta señal pertenezca a la clase c_i. Una vez obtenidos todos los scores, se forma un vector $S = \{s_i^j\}$ con todos ellos. Este vector de scores puede ser utilizado como un nuevo vector de parámetros que será clasificado mediante la SVM de fusión, dando así el resultado final. Con un entrenamiento adecuado, se espera que la SVM de fusión sea capaz de aprender la distribución de scores de los aciertos y errores de los clasificadores q_j, permitiendo así mejorar los resultados individuales de cada sistema.

Para las pruebas descritas en este trabajo, el sistema de fusión se ha entrenado utilizando los scores de las pruebas de desarrollo. Aunque en principio las SVM no proporcionan scores, sino decisiones finales, se ha utilizado la técnica descrita por Wu *et al.* (2004) para la estimación de probabilidades a partir de los valores de la función de decisión. Gracias a esta técnica, es posible calcular la probabilidad de que la muestra a clasificar pertenezca a cada una de las clases, partiendo de los valores de la función de decisión de una SVM. Estos valores de probabilidad se han utilizado como scores para el sistema de fusión.

Los resultados obtenidos con la fusión tardía se muestran en la Tabla 4.8. La combinación de los parámetros supra-segmentales (columna 1) proporciona resultados muy similares tanto con fusión temprana (columna 5 de la Tabla 4.7) como tardía. Por el contrario, la combinación de parámetros segmentales presenta una gran mejora al utilizar una fusión de scores (columna 2). Sin embargo, esta mejora es en parte debida a la combinación de los dos flujos (sonoro y sordo) de los LFPC y las primitivas de prosodia, lo cual no es posible con la fusión temprana.

El modelar los flujos de tramas sonoras y sordas por separado y combinarlas posteriormente mediante una fusión tardía proporciona una mejora significativa. Mientras que con la parametrización LFPC se alcanza una precisión de 69,9 % con 20 parámetros y 72,2 % con todos los 54 (columna 6 de la Tabla 4.7), modelar

por separado sus flujos y fusionarlos posteriormente (columna 3) proporciona un 72 % y 76,5 % respectivamente, lo que representa una reducción del error del 7 % y el 15 % respectivamente. La razón de esta mejora puede ser el hecho de que las tramas sordas y sonoras tienen características espectrales muy diferentes. Al modelarlas mediante un único GMM, el modelo obtenido es menos robusto debido a estas diferencias. Sin embargo, por separado, cada modelo captura adecuadamente las características de cada uno de los flujos. La fusión tardía permite combinar estas informaciones sin perder precisión. También se puede observar una mejora apreciable al combinar los flujos de las primitivas de prosodia (columna 4). En vista de estos resultados, parece aconsejable mantener los dos flujos por separado.

Como se ha indicado, la fusión tardía también puede ser utilizada para combinar sistemas basados en parámetros segmentales y supra-segmentales. La combinación de los resultados de la prosodia con LFPC (columna 5) proporciona una precisión similar a la obtenida al combinar el sistema de estadísticos espectrales a largo plazo con las primitivas de prosodia a nivel de trama (columna 6). En ambos casos la precisión es mayor que cuando se utilizan únicamente los estadísticos de espectro (el mejor de los sistemas aislados) y ligeramente mejor que usando la fusión temprana de todas las parametrizaciones (el mejor de los sistemas de fusión temprana).

Cuando se combinan los LFPC con los estadísticos de espectro (columna 7), o la prosodia con sus primitivas (columna 8), también hay una mejora significativa. Esto significa que la fusión de sistemas que utilizan parámetros extraídos de la misma fuente de información acústica, pero con un intervalo de integración diferente, también permite reducir los errores de la clasificación.

Como último experimento se ha probado a combinar todas las parametrizaciones con el sistema de fusión tardía (columna 9): estadísticas de espectro, prosodia y calidad de voz junto con las primitivas de prosodia y LFPC calculadas para cada trama, con flujos sonoros y sordos separados. Esto da un total de 7 clasificadores combinados. Los resultados obtenidos en esta última prueba están entre los mejores de todos los sistemas analizados: Un 78,3 % de acierto con todos los parámetros y un 76,8 % cuando se utilizan parámetros seleccionados. Esto representa un 20 % de reducción de error con respecto al mejor sistema de fusión temprana.

TABLA 4.7: *Precisión de las pruebas de desarrollo y finales sobre la base de datos* Berlin*. Se muestran los resultados con todos los parámetros y con el número óptimo de parámetros seleccionado durante el desarrollo. Valores de precisión en porcentaje.*

	Prosodia	**VQ**	**Espectral**	**Prosodia+ VQ**	**Prosodia+ VQ+ Espectral**	**LFPC**	**LFPC sonora**	**LFPC sorda**	**PP sonora**	**PP sorda**	**LFPC+ PP sonora**	**LFPC+ PP sorda**
# param.	39	5	96	17	152	20	20	20	6	3	20	20
Desarrollo	65,5	49,4	75,4	67,4	77,9	70,9	67,7	62,8	65,3	50,8	68,2	59,7
Final	62,6	47,3	70,5	60,9	72,2	69,9	63,2	61,5	60,8	48,4	66,4	57,0
# param.	54	5	324	59	383	54	54	54	6	3	60	57
Desarrollo	64,5	49,4	74,3	66,7	77,6	72,9	71,2	64,3	65,3	50,8	75,1	65,8
Final	64,5	47,3	70,7	63,6	72,5	72,2	70,5	64,7	60,8	48,4	66,5	63,6

TABLA 4.8: *Precisión de las pruebas de fusión tardía sobre la base de datos* Berlin*. Se muestran los resultados utilizando todos los parámetros y en el caso de utilizar el número óptimo de parámetros seleccionado durante el desarrollo. Valores de precisión en porcentaje.*

	Prosodia+ VQ+ Espectral	**LFPC-V+ LFPC-UV+ PP-V+PP-UV**	**LFPC-V+ LFPC-UV**	**PP-V+ PP-UV**	**Prosodia+ LFPC-V+ LFPC-UV**	**Espectral+ PP-V+ PP-UV**	**Espectral+ LFPC-V+ LFPC-UV**	**Prosodia+ PP-V+ PP-UV**	**TODOS**
Param. sel.	71,8	74,4	72,0	65,4	75,5	74,8	76,6	69,2	76,8
Todos param.	72,2	77,4	76,5	65,4	76,6	74,4	76,6	69,0	78,3

4.6. Conclusiones

En este capítulo se ha comparado la capacidad de discriminación proporcionada por diferentes parametrizaciones utilizadas tradicionalmente para la identificación automática de emociones. Para ello se han realizado varias medidas de discriminalidad, mediante el cálculo de las dispersiones intra e inter emoción, agrupación ciega no supervisada y selección de parámetros. Con el objetivo de verificar los resultados obtenidos, también se han realizado experimentos empíricos de identificación automática de emociones independiente de locutor.

La mayoría de los estudios publicados que se centran en el análisis de las características de la voz en función del estado emocional del locutor muestran que, considerados individualmente, los cambios en los parámetros prosódicos son más significativos que los cambios en la envolvente espectral. Sin embargo, los resultados obtenidos en este capítulo sugieren que, si se considera el conjunto completo de parámetros, las parametrizaciones espectrales proporcionan más información que las prosódicas.

Las medidas basadas en las dispersiones intra e inter emoción proporcionan un mecanismo para estimar el grado de solapamiento entre las clases usando todo el conjunto de parámetros en lugar de valores individuales. Estas medidas se han complementado con una agrupación ciega no supervisada para comprobar si estas parametrizaciones proporcionan una correcta separación de las emociones. Ambos métodos revelan que los estadísticos supra-segmentales de la envolvente espectral proporcionan mayor separación entre las emociones que los parámetros prosódicos tradicionales. Esto permite explicar los resultados de las pruebas experimentales, donde los estadísticos espectrales consiguen resultados significativamente mejores. Las medidas también muestran que la combinación de parametrizaciones supra-segmentales de la envolvente espectral, la prosodia y la calidad de voz reduce el solapamiento entre emociones, confirmando que el uso de parámetros extraídos a partir de diferentes fuentes de información permite reducir el error de identificación. Esta misma conclusión se puede extraer de los resultados experimentales. Sin embargo, parece ser cierta únicamente para las parametrizaciones supra-segmentales. La fusión temprana de LFPC y primitivas de prosodia, tanto para el flujo de tramas sonoras como el de tramas sordas, no alcanza el nivel de precisión logrado al utilizar únicamente los LFPC del mismo flujo.

Por desgracia, los cálculos basados en dispersiones o en agrupación ciega no son adecuados para el análisis de parametrizaciones a nivel de segmento. En su lugar, se ha aplicado un algoritmo de ranking de parámetros para detectar aquellos que son más discriminantes. Aunque muchos trabajos publicados en este campo mencionan el uso de algún tipo de algoritmo de selección de parámetros, son pocos los que comentan los resultados de esta selección. En los casos en los que se proporciona este resultado, los estadísticos a largo plazo de la envolvente es-

pectral suelen ser seleccionados antes que los prosódicos (Schuller *et al.*, 2005; Vogt y André, 2006), sugiriendo que estos parámetros espectrales proporcionan mayor información acerca del estado emocional del locutor. Los resultados del ranking presentados en este capítulo coinciden con esta conclusión. Sin embargo, en el caso de los parámetros segmentales, las características relativas a las primitivas de prosodia (concretamente las derivados de la curva de entonación) son clasificadas en las primeras posiciones. Por lo tanto, se puede concluir que, al menos cuando se consideran de forma individual, proporcionan una gran cantidad de información. Otra vez, los resultados experimentales parecen confirmar esta conclusión, ya que la precisión obtenida por el flujo sonoro de las primitivas de prosodia alcanza valores superiores a los obtenidos mediante el mismo número de parámetros LFPC para el mismo flujo (Figura 4.5) y similares a los obtenidos mediante LFPC sin separación de flujos. Sin embargo, considerando las parametrizaciones completas, los LFPC sin separación de flujos consiguen resultados significativamente superiores.

Los resultados de los experimentos de identificación automática presentados permiten analizar la precisión final, no sólo de cada parametrización, sino también de sus combinaciones. De esta forma, es posible comprobar la mejora obtenida a través de estas combinaciones, y determinar si la combinación es ventajosa o no. Por ejemplo, aunque las primitivas de prosodia alcanzan los primeros puestos en el ranking de parámetros, y de forma individual logran una precisión considerable (teniendo en cuenta el reducido número de parámetros), se ha comprobado que la combinación de parámetros LFPC a nivel de trama y estas primitivas no mejora significativamente los resultados obtenidos usando sólo los LFPC. Únicamente cuando se utiliza un reducido número de parámetros esta combinación resulta provechosa (Figura 4.5). También en el caso de las parametrizaciones supra-segmentales, las características prosódicas ofrecen mejores resultados que los estadísticos de espectro sólo cuando se utilizan pocos parámetros (Figura 4.4a). Según estos resultados, se puede decir que las características prosódicas tradicionales parecen ser las más apropiadas para la identificación de emociones sólo cuando se consideran de forma individual o en un conjunto muy reducido de parámetros. En el caso de utilizar un mayor número de parámetros, las características prosódicas proporcionan mejores resultados.

Los resultados de los experimentos llevados a cabo mediante fusión tardía sugieren que el uso combinado de parametrizaciones extraídas de una misma fuente de información pero con una escala de tiempo diferente (parámetros segmentales y supra-segmentales) mejora la precisión del sistema. Las diferencias en la escala de tiempo y en el clasificador utilizado hacen que cada subsistema retenga diferentes características de la emoción. De hecho, uno de los mejores resultados se ha conseguido con la fusión tardía de estadísticos a largo plazo de LFPC con los flujos de tramas sonoras y sordas de LFPC a nivel de trama. Cuando se utilizan

los parámetros seleccionados, este sistema sólo ha sido superado por la fusión tardía de todos los parámetros considerados. Sin embargo, este último sistema es mucho más complejo, mientras que la mejora obtenida es muy pequeña. El uso de todos los parámetros requiere estimar los LFPC, valores de entonación, decisiones VUV, entonación, marcado a período de pitch, filtrado inverso y detección de vocales. Esto complica la etapa de parametrización, haciéndola computacionalmente costosa. Además, utiliza siete clasificadores diferentes al que hay que añadir el sistema de fusión. Por el contrario, el sistema espectral sólo requiere del cálculo de LFPC, algunos estadísticos sencillos de calcular y la decisión VUV para poder separar los flujos, reduciendo el número de clasificadores a tres más el de fusión. La diferencia en la precisión puede no justificar el incrementar la complejidad del sistema hasta tal grado.

A modo de resumen, la siguiente lista agrupa las conclusiones extraídas:

- Las características extraídas de las curvas de prosodia (tanto a nivel segmental como supra-segmental) son más significativas para la identificación de emociones si se consideran individualmente o en grupos reducidos.

- En caso de considerar un mayor número de parámetros, las características espectrales proporcionan una mayor separación de las emociones.

- Los parámetros de calidad de voz proporcionan una discriminación muy pobre. Probablemente este resultado se deba a errores durante la parametrización automática. En caso de querer utilizar características de calidad de voz, será necesario realizar una revisión manual de los parámetros o mejorar los sistemas de caracterización.

- En las parametrizaciones segmentales, separar los flujos sonoro y sordo y combinar los resultados mediante fusión tardía proporciona en general una mejora en los resultados.

- Para el caso de las características supra-segmentales, la combinación de parametrizaciones espectrales, prosódicas y de calidad de voz proporciona un incremento de la precisión del sistema con respecto a las parametrizaciones individuales. Esta mejora es comparable tanto en el caso de la fusión temprana como en la fusión tardía.

- Sin embargo, la combinación de las primitivas de prosodia con la parametrización espectral a nivel de segmento (separando flujos sonoro y sordo) mediante fusión temprana proporciona peores resultados que si se utiliza sólo el espectro. Combinarlos mediante fusión tardía proporciona mejoras significativas.

- Utilizar la fusión tardía para combinar información extraída de una misma fuente pero con intervalos de integración diferentes (a nivel de trama y a nivel de frase) aumenta la precisión del sistema con respecto a utilizar una única base de tiempos.
- Con todo esto, la fusión tardía de estadísticos de espectro a largo plazo con los flujos sonoro y sordo de la parametrización espectral a nivel de trama proporciona uno de los mejores resultados. Añadir a este sistema las primitivas de prosodia y los estadísticos prosódicos sólo proporciona una mínima mejora.

Capítulo 5

Validación de resultados con emociones naturales

Índice

EL análisis presentado en el capítulo 4 se ha realizado sobre una base de datos de emociones actuadas. Gracias al procedimiento utilizado durante la creación de esta base de datos, y en particular a la evaluación perceptual (ver sección 4.1), se puede considerar que contiene grabaciones de alta naturalidad. Por lo tanto, se espera que las conclusiones obtenidas en el capítulo anterior sean válidas también para emociones reales no actuadas. Sin embargo, es necesario comprobar si esta suposición es cierta.

Para ello se han realizado una serie de pruebas utilizando la base de datos *AIBO*. Se trata de una base de datos muy realista, con emociones naturales y habla espontánea, lo que permite comprobar si las conclusiones extraídas durante la etapa de análisis son generalizables a situaciones reales, o sólo debidas a la naturaleza actuada de las señales utilizadas. Además, el conjunto de emociones que proporciona es diferente, por lo que estas pruebas también permiten verificar si los resultados obtenidos son válidos para otras emociones no consideradas durante el análisis anterior.

5.1. Descripción de la base de datos de trabajo: *Aibo*

La base de datos *AIBO* (Batliner *et al.*, 2006) contiene cerca de 18.000 grabaciones, realizadas a 21 niños y 30 niñas de entre 10 y 13 años, mientras jugaban con el robot Sony-AIBO[1]. Se pidió a estos niños que dirigieran al robot mediante comandos de voz a través de un circuito predefinido, aunque en realidad era controlado por un operario desde otra sala. Puesto que los niños creían que el robot obedecía sus órdenes, tenían la impresión de que a veces les hacía caso y otras no. Los comandos y reacciones vocales de los niños fueron grabadas y forman la base de datos. Se trata, por tanto, de una base de datos de habla emocionada infantil espontánea, obtenida mediante la técnica de WOZ.

[1] http://support.sony-europe.com/aibo/

5.1.1. Etiquetado de las grabaciones

Una vez finalizado el proceso de grabación, cinco lingüistas etiquetaron las grabaciones a nivel de palabra, según once categorías emocionales: aburrimiento, desesperación, enfado, enfático, felicidad, irritación, neutro, maternal, recriminativo, sorpresa y otro. La etiqueta final de cada palabra se seleccionó mediante voto por mayoría. En total se etiquetaron 43.694 palabras.

Las once categorías consideradas inicialmente son excesivamente específicas, y provocan además un gran desequilibrio en el número de ejemplos de cada clase. Algunas categorías quedan prácticamente vacías (aburrimiento: 11 palabras; desesperación: 3 palabras; sorpresa: 0 palabras). Teniendo en cuenta que varias de las emociones son similares entre sí, a la hora de utilizar la base de datos estas etiquetas se suelen reagrupar en cuatro o cinco emociones más generales. Por ejemplo, para la iniciativa CEICES de la red de excelencia HUMAINE de la Unión Europea (Schuller *et al.*, 2008), se utilizaron cuatro etiquetas: enfado (agrupando enfado, irritación y recriminativo), enfático, maternal y neutro. El resto de clases fueron descartadas por disponer de pocos ejemplos. Sin embargo, para las pruebas de *Emotion Challenge* (Schuller *et al.*, 2009) se utilizaron cinco: enfado (agrupando enfado, irritación y recriminativo), enfático, neutro, positivo (agrupando felicidad y maternal) y resto (agrupando todas las demás). En este trabajo se ha utilizado esta última clasificación.

Además del etiquetado a nivel de palabra, esta base de datos proporciona también una etiqueta para cada grabación. La emoción correspondiente a cada grabación se deduce a partir de las etiquetas que los evaluadores han asignado a cada palabra de esa frase, utilizando un método heurístico (Batliner *et al.*, 2006). Calculando la relación entre el número de etiquetas de cada emoción, y siguiendo unas reglas predefinidas, el método estima la etiqueta final de la grabación. Debido a la forma en la que está definido el algoritmo heurístico, *siempre* asigna una etiqueta de emoción a cada frase, aunque ésta contenga palabras con diferente etiquetado, o aunque los etiquetadores no hayan llegado a un acuerdo. Junto con la etiqueta estimada por el algoritmo también se proporciona la tasa de votos correspondientes a esa emoción que los etiquetadores han asignado a las palabras de dicha frase. Supongamos que N es el número de palabras de una frase y que c_i^j es la etiqueta adjudicada por el etiquetador i a la palabra j ($i = 1 \ldots 5$; $j = 1 \ldots N$). Supongamos asimismo que C es la etiqueta final asignada por el método heurístico. La tasa de votos proporcionada se calcula como:

$$S = \frac{\sum_{j=1}^{N} \sum_{i=1}^{5} f(c_i^j)}{5N} \qquad f(c_i^j) = \begin{cases} 1 & \text{si } c_i^j = C \\ 0 & \text{si } c_i^j \neq C \end{cases} \tag{5.1}$$

Este valor puede considerarse un indicador del consenso de los etiquetadores a la hora de describir las emociones detectadas dentro de la grabación. Valores altos de *S* indican que la mayoría de las palabras han sido etiquetadas como *C* por la mayoría de los etiquetadores, mientras que valores bajos indican que no está realmente claro cuál es la emoción de la frase completa. Esto puede ocurrir si el contenido emocional no es claramente identificable o si aparece más de una emoción en una misma grabación.

La Tabla 5.1 presenta un breve resumen de la distribución de los valores de estos indicadores para las diferentes emociones que componen la base de datos. Se puede comprobar que en más de 5.000 frases (el 27,6 % de toda la base de datos) estos indicadores tienen un valor menor de 0,5. Es decir, que de todos los votos emitidos para esa frase, menos de la mitad coinciden con la etiqueta final impuesta por el método heurístico. Es más, unas 1.200 frases (el 6,6 %) tienen el indicador con valor cero, reflejando que ni uno solo de los votos de los etiquetadores hace referencia a la emoción final. Todas estas frases con indicador de valor cero se corresponden a la emoción *resto*. Esto es un efecto directo de la forma en la que está definido el algoritmo heurístico, ya que en el caso de que no haya consenso entre las etiquetas, tiene una fuerte tendencia a asignar la frase a esta clase *resto*, aunque no haya ni una sola etiqueta de este valor.

TABLA 5.1: *Distribución de los indicadores de consenso en el etiquetado de la base de datos* AIBO. *Los valores reflejan número de frases.*

Etiqueta	$0-0,25$	$0,25-0,5$	$0,5-0,75$	$0,75-1$
Enfado	2	774	433	283
Enfático	0	2600	799	202
Neutro	0	0	1861	9106
Positivo	0	398	323	168
Resto	1207	46	14	0
TOTAL	1209	3818	3430	9759

Se ha considerado que una frase con menos del 50 % de votos coincidentes tiene una emoción incierta, ya que ni siquiera los etiquetadores humanos han podido llegar a un acuerdo sobre qué emoción representa. Si estas frases se utilizan durante el entrenamiento de los sistemas, sólo proporcionarán ruido, dando lugar a modelos poco robustos. Si se utilizan durante la etapa de pruebas, es muy probable que la etiqueta asignada por el clasificador automático para estas frases no coincida con la estimada por el método heurístico, contabilizándose como errores de identificación. Sin embargo, se trataría de falsos errores, pues la etiqueta emo-

cional proporcionada en la base de datos tampoco es un indicador fiable que se pueda tomar como referencia.

Por lo tanto, se ha decidido eliminar de la base de datos todas las grabaciones con indicadores inferiores a 0,5. Después de este proceso de limpieza, sólo quedan 14 señales etiquetadas como *resto*, lo cual es insuficientes para realizar pruebas significativas. En consecuencia, se ha tenido que suprimir esta clase emocional, dejando la base de datos con un total de 13.174 frases y cuatro categorías emocionales: *enfado*, *enfático*, *neutro* y *positivo*.

5.1.2. División de la base de datos

Las grabaciones se realizaron en dos colegios diferentes, *Ohm* (26 niños y niñas) y *Mont* (25 niños y niñas). Aprovechando esta circunstancia, los autores de la base de datos sugieren utilizar las señales correspondientes al colegio *Ohm* para realizar el entrenamiento del sistema y las correspondientes al colegio *Mont* para las pruebas finales. Para el estudio presentado en este trabajo se ha mantenido este convenio, ya que de esta forma se asegura la independencia de locutor en las pruebas. La parte de entrenamiento se utilizará para realizar todos los procesos de análisis de los parámetros y el desarrollo de los sistemas de clasificación, reservando la parte de pruebas para los experimentos finales de identificación automática. La Tabla 5.2 resume el contenido de la base de datos, tanto para la parte de entrenamiento como para la de pruebas. Puede comprobarse que el número de ejemplos disponible para cada emoción está fuertemente desequilibrado, estando la mayoría de las señales etiquetadas como neutras.

TABLA 5.2: *Distribución de las señales en la base de datos* AIBO *para cada una de las emociones consideradas.*

	Enfado	**Enfático**	**Neutro**	**Positivo**	**Total**
Entrenam.	424 (6,0%)	630 (9,0%)	5589 (79,4%)	398 (5,7%)	7041
Pruebas	292 (4,8%)	371 (6,1%)	5377 (87,7%)	93 (1,5%)	6133

Al igual que en las pruebas realizadas sobre la base de datos *Berlin*, es necesario disponer de resultados de desarrollo para la optimización de los modelos y el entrenamiento del sistema de fusión tardía. Para obtener estos resultados de desarrollo, los locutores del conjunto de entrenamiento se han repartido aleatoriamente en cinco grupos de cinco locutores cada uno, sobre los que se ha aplicado

una validación cruzada. El locutor con menos señales se eliminó de las pruebas de desarrollo para mantener el equilibrio de locutores entre los bloques. Tal y como muestra la Tabla 5.3, la distribución de las emociones en estos bloques es aproximadamente la misma que en la base de datos completa.

TABLA 5.3: *Distribución de las señales en los bloques de desarrollo definidos en la base de datos* AIBO*, para cada una de las emociones consideradas.*

	Enfado	**Enfático**	**Neutro**	**Positivo**	**Total**
Bloque 1	64 (4,8%)	190 (14,1%)	1059 (78,6%)	35 (2,6%)	1348
Bloque 2	90 (6,8%)	88 (6,6%)	976 (73,5%)	174 (13,1%)	1328
Bloque 3	95 (6,3%)	73 (4,9%)	1272 (84,7%)	63 (4,2%)	1503
Bloque 4	108 (7,9%)	111 (8,1%)	1076 (79,0%)	68 (5,0%)	1363
Bloque 5	64 (4,7%)	152 (11,2%)	1085 (79,8%)	58 (4,3%)	1359
Total	421 (6,1%)	614 (8,9%)	5468 (79,2%)	398 (5,8%)	6901

5.2. Medidas de variabilidad en emociones naturales

La Tabla 5.4 muestra las medidas de discriminalidad de las parametrizaciones supra-segmentales en la base de datos *AIBO*, calculadas según la expresión (4.4). Como era de esperar, los valores de discriminalidad son mucho menores que para el caso de las emociones actuadas. A ello contribuyen dos factores. Por un lado, las emociones naturales suelen darse con mucha menos intensidad que las actuadas, lo que reduce la dispersión inter-clase, haciendo que sea más difícil el distinguirlas. Por otro, al tratarse de habla espontánea, la calidad de las locuciones es en algunos casos muy pobre, presentando ruidos y distorsiones y una pronunciación a menudo poco cuidada. Esto hace que la estimación de los parámetros sea poco robusta, provocando un incremento en la dispersión intra-clase, y por lo tanto, una mayor confusión entre las emociones.

TABLA 5.4: *Discriminalidad de los parámetros supra-segmentales en* AIBO.

Parámetros	J_1
Prosódicos	0,66
Calidad	0,14
Espectrales	2,35
Pros.+Calidad	0,74
Pros.+Calidad+Espec.	3,00

Sin embargo, las conclusiones generales obtenidas para el caso de las emociones actuadas siguen siendo válidas:

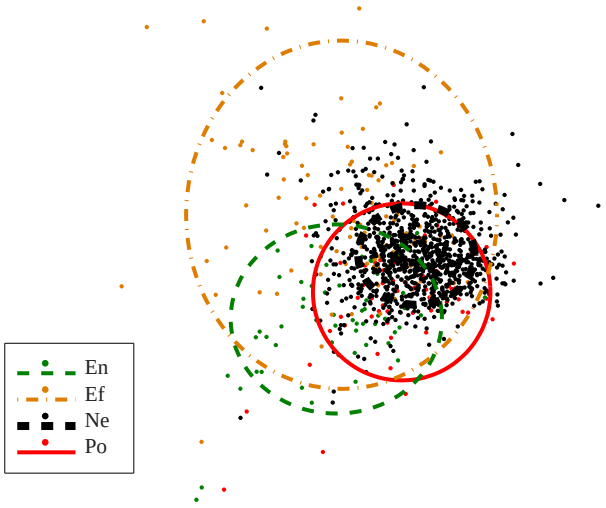
- Los estadísticos de envolvente espectral siguen proporcionando mayor separación entre emociones que la parametrización prosódica.
- Los valores de calidad de voz no permiten la correcta identificación de las emociones, aunque cuando se combinan con la prosodia, permiten mejorar la separación hasta cierto punto.
- La combinación de todos los parámetros supra-segmentales proporciona la máxima separación entre las clases.

La Figura 5.1 muestra el diagrama de dispersión de los estadísticos de espectro y de prosodia proyectados sobre las dos direcciones más discriminantes obtenidas según un análisis LDA. Puede comprobarse que las características espectrales realmente proporcionan una mayor separación entre las emociones, al igual que ocurría en la base de datos actuada *Berlin*.

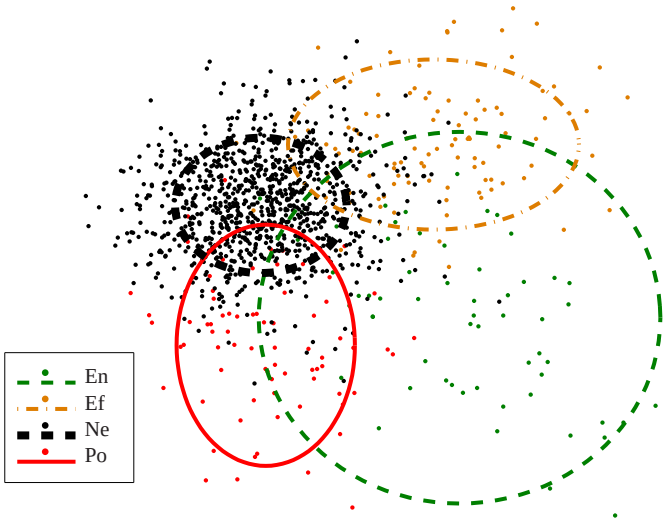
5.3. Selección de parámetros en emociones naturales

Se ha aplicado el algoritmo mRMR descrito en la sección 4.4 a las señales de entrenamiento de la base de datos, obteniendo un nuevo ranking para los parámetros considerados, así como para sus diferentes combinaciones de fusión temprana. La Tabla 5.5 muestra los 30 primeros parámetros seleccionados en el caso de las parametrizaciones combinadas, lo que permite comparar los resultados de esta selección con el ranking obtenido para las emociones actuadas (Tabla 4.6).

Era de esperar que el orden exacto de los parámetros fuera diferente en ambas bases de datos. Si dos parámetros proporcionan una información similar, pueden



(A) *Dispersión de las clases con parámetros prosódicos.*



(B) *Dispersión de las clases con parámetros espectrales.*

FIGURA 5.1: *Diagrama de dispersión de los parámetros supra-segmentales prosódicos (a) y espectrales (b) para la base de datos* AIBO, *proyectados sobre las dos direcciones más discriminantes según el análisis LDA. Las elipses delimitan la región a dos desviaciones estándar del centroide.* En*: enfado,* Ef*: enfático,* Ne*: neutro,* Po*: positivo.*

TABLA 5.5: *Los primeros 30 parámetros seleccionados en las parametrizaciones combinadas para la base de datos* AIBO. VQ*: calidad de voz,* PP*: primitivas de prosodia.*

#	Parámetros supra-segmentales		Parámetros segmentales	
	Prosodia+ VQ	**Prosodia+ VQ+ Espectral**	**LFPC+ PP sonoros**	**LFPC+ PP sordos**
1	$\mathrm{R}(Pow)$	$\sigma^2(LFPC_{18})$	ΔPow	$\Delta^2 Pow$
2	$NLvPcn$	$\mathrm{K}(LFPC_9)$	ΔF_0	$LFPC_{16}$
3	NAQ	$\mathrm{Sk}(LFPC_2)$	$\Delta^2 Pow$	$\Delta LFPC_{15}$
4	$\mathrm{K}(\Delta^2 Pow)$	$\mathrm{E}(LFPC_1)$	$LFPC_{17}$	$LFPC_3$
5	$\mathrm{E}(\Delta^2 F_0)$	$\min(\Delta LFPC_3)$	$\Delta^2 F_0$	$\Delta^2 LFPC_{17}$
6	$\mathrm{E}(Vdur)$	$\mathrm{K}(\Delta^2 LFPC_{12})$	$\Delta^2 LFPC_6$	$\Delta LFPC_3$
7	SB	NAQ	$\Delta^2 LFPC_{11}$	$\Delta^2 LFPC_7$
8	$\sigma^2(\Delta^2 Pow)$	$LvPsl$	$\Delta^2 LFPC_4$	$\Delta LFPC_{18}$
9	$\min(F_0)$	$\mathrm{Sk}(\Delta LFPC_4)$	$\Delta LFPC_{17}$	$\Delta^2 LFPC_{14}$
10	$LvPsl$	$\mathrm{E}(Vdur)$	$\Delta^2 LFPC_5$	$LFPC_1$
11	$\mathrm{Sk}(Pow)$	$NLvPcn$	$\Delta^2 LFPC_{16}$	$LFPC_5$
12	LvF_0sl	$\mathrm{E}(\Delta LFPC_{16})$	$\Delta LFPC_5$	$\Delta^2 LFPC_{18}$
13	$\mathrm{Sk}(\Delta Pow)$	$\sigma^2(\Delta^2 F_0)$	$\Delta^2 LFPC_7$	$\Delta^2 LFPC_3$
14	Jit	$\min(\Delta^2 LFPC_6)$	F_0	$\Delta LFPC_{16}$
15	$LvVdur$	SB	$\Delta^2 LFPC_3$	$\Delta^2 LFPC_{11}$
16	$\sigma^2(\Delta F_0)$	$\mathrm{Sk}(\Delta LFPC_8)$	$\Delta LFPC_6$	$\Delta LFPC_{14}$
17	$\mathrm{E}(Psl)$	LvF_0sl	$\Delta LFPC_2$	$LFPC_8$
18	$LvPcn$	$\min(F_0)$	$\Delta^2 LFPC_9$	$\Delta LFPC_1$
19	$\mathrm{E}(F_0)$	$\mathrm{Sk}(LFPC_7)$	$\Delta LFPC_7$	$\Delta^2 LFPC_{16}$
20	$\min(\Delta Pow)$	$\mathrm{K}(LFPC_1)$	$\Delta^2 LFPC_{13}$	$\Delta LFPC_{11}$
21	$\mathrm{K}(\Delta Pow)$	$\mathrm{Sk}(LFPC_{14})$	$\Delta LFPC_4$	$LFPC_2$
22	$\sigma^2(Pow)$	$\mathrm{E}(LFPC_{18})$	$\Delta^2 LFPC_2$	$\Delta^2 LFPC_{12}$
23	$\mathrm{K}(F_0)$	$\mathrm{Sk}(\Delta LFPC_1)$	$\Delta LFPC_{16}$	$\Delta LFPC_{17}$
24	$\sigma^2(Vdur)$	$\mathrm{Sk}(\Delta^2 LFPC_9)$	$\Delta^2 LFPC_8$	$\Delta^2 LFPC_{15}$
25	$\mathrm{E}(\Delta F_0)$	$\mathrm{K}(\Delta LFPC_{15})$	$\Delta^2 LFPC_{15}$	$\Delta^2 LFPC_5$
26	$\min(Pow)$	$\min(\Delta LFPC_4)$	$\Delta LFPC_{11}$	$LFPC_4$
27	$\min(\Delta^2 Pow)$	$\mathrm{E}(\Delta^2 F_0)$	$LFPC_3$	$\Delta LFPC_{13}$
28	TLT	$\mathrm{R}(LFPC_1)$	$\Delta^2 LFPC_{17}$	$\Delta^2 LFPC_{13}$
29	LvF_0cn	$\mathrm{K}(LFPC_6)$	$\Delta^2 LFPC_{12}$	$\Delta LFPC_5$
30	$\mathrm{Sk}(\Delta^2 Pow)$	$NLvVdur$	$\Delta LFPC_{14}$	$\Delta^2 LFPC_1$

considerarse redundantes, y no es relevante cuál de los dos se selecciona primero. Por lo tanto, no estamos interesados en comparar esta ordenación posición a posición, sino en analizar la naturaleza de los parámetros que ocupan los primeros puestos.

Por ejemplo, en la combinación de todos los parámetros supra-segmentales, los estadísticos de envolvente espectral siguen ocupando las primeras posiciones, aunque las características de calidad de voz se encuentran algo mejor situadas. Entre los primeros 20 parámetros encontramos 12 espectrales, 6 prosódicos y dos de calidad de voz, mientras que en el caso de la base de datos *Berlin* la calidad de voz no figuraba hasta la posición 23. Por tanto, parece que los valores asociados a la señal glotal tienen una importancia relativa algo mayor cuando se trata de emociones naturales y habla espontánea.

Las diferencias más importantes entre las dos tablas se encuentran en los parámetros a nivel de segmento. En el caso de los flujos sonoros, cuatro de las cinco primeras posiciones están ocupadas por valores relativos a las primitivas de prosodia. Esto podría llevar a pensar que estas primitivas de prosodia transportan la mayor información acerca de la emoción. Sin embargo, y como se podrá comprobar en los resultados de los experimentos de identificación automática, llegar a esta conclusión a partir de una selección que considera los parámetros de forma individual puede ser engañoso. En realidad sólo se puede deducir que estos parámetros son muy informativos cuando se consideran de forma individual. Pero no se puede generalizar y suponer que su combinación es la mejor de todas las posibles. Esta misma precaución es aplicable a las características de calidad de voz en el caso de los parámetros supra-segmentales.

5.4. Experimentos de identificación de emociones naturales

La capacidad de las diferentes parametrizaciones a la hora de discriminar emociones naturales también se ha comprobado experimentalmente mediante pruebas de identificación automática sobre la base de datos *AIBO*. Puesto que esta base de datos está fuertemente desequilibrada en el número de grabaciones disponibles para cada emoción, se ha utilizado la UAR como medida de precisión de los experimentos, tal y como ya se ha hecho anteriormente con la base de datos *Berlin*.

Toda la arquitectura de los sistemas de identificación automática, así como los clasificadores utilizados, son los mismos que los descritos en la sección 4.5.1 para la anterior base de datos. La única excepción es la división de la base de datos para generar resultados de desarrollo, para lo cual se ha aplicado el mecanismo descrito en la sección 5.1.2. Al igual que en el caso de las emociones actuadas, estas prue-

bas de desarrollo se han utilizado para estimar los valores óptimos del número de mezclas de los GMM, la dispersión del kernel RBF, el coste de clasificación errónea de la SVM y el número óptimo de parámetros.

5.4.1. Selección del número de parámetros

El número óptimo de parámetros se ha determinado repitiendo las pruebas de desarrollo con un parámetro más cada vez, según el ranking de la Tabla 5.5. La Figura 5.2 muestra la evolución de la precisión estimada para las características supra-segmentales según estas pruebas de desarrollo. Aparentemente la precisión del sistema se satura mucho antes que en el caso de las emociones actuadas. Mientras que en la base de datos *Berlin* la precisión se estabiliza alrededor de 20 parámetros, en el caso de *AIBO* la saturación llega con 10 parámetros. Sin embargo, la precisión mantiene una ligera tendencia positiva, lo que hace que el número óptimo de parámetros (aquél con el que se consigue la mínima tasa de error) sea aproximadamente igual en ambos casos.

Según la Figura 5.2(a), parece que los estadísticos de envolvente espectral proporcionan menor tasa de acierto que los parámetros prosódicos. Resulta curioso, puesto que en el ranking de parámetros los valores espectrales se encuentran en mejores posiciones que los prosódicos. Sin embargo, si se observa el comportamiento del sistema espectral a medida que se incrementa el número de parámetros (Figura 5.2(b)), puede comprobarse que alcanza un 56,1 % de precisión con 119 parámetros, un valor muy próximo al obtenido con la prosodia: 55,7 % con 25 parámetros. El aumento del número de parámetros permite a los estadísticos espectrales compensar el incremento de información proporcionado por la prosodia.

Los parámetros de calidad de voz son los menos adecuados para la identificación de emociones naturales, alcanzando un 41,6 % de precisión máxima. Evidentemente, el reducido número de parámetros (tan sólo cinco) supone una desventaja para esta parametrización. Sin embargo, utilizando sólo los cinco mejores parámetros prosódicos se alcanza un 49,8 %, una diferencia muy significativa. En general, cualquier parametrización proporciona mejores resultados que la calidad de voz, para un mismo número de parámetros. Además, la combinación de la prosodia y la calidad de voz no aporta ventajas con respecto a utilizar la prosodia de forma aislada.

En vista de estos resultados, se han analizado los valores de calidad de voz resultantes de la parametrización, y se ha comprobado que estos valores tienen una gran dispersión intra-emoción. Se cree que este efecto, que ya podía intuirse por los valores de J_1 (ver Tabla 5.4), es debido a la naturaleza espontánea de las locuciones. Los valores de calidad de voz se calculan a partir de un proceso de filtrado inverso, y este filtrado no es preciso si la señal de voz no es limpia y estable. La voz espontánea no tiene una pronunciación ni una vocalización cuidada,

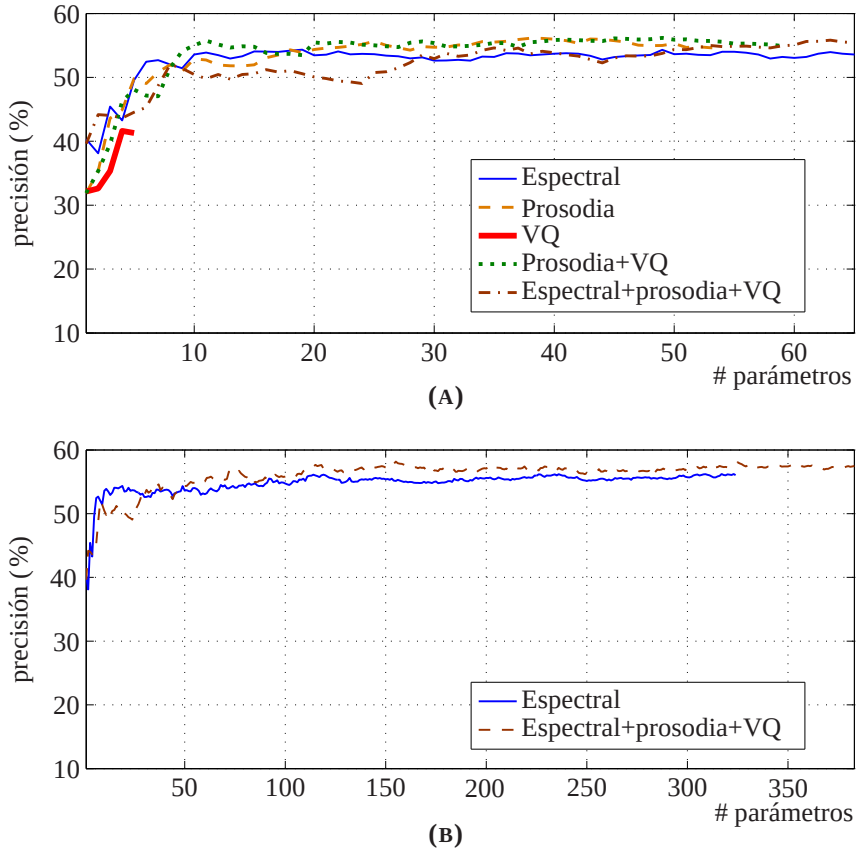


FIGURA 5.2: *Resultados de las pruebas de desarrollo para las parametrizaciones supra-segmentales en la base de datos AIBO. Los gráficos muestran la precisión del sistema en función del número de parámetros. (a) Comparativa entre todas las parametrizaciones analizadas. (b) Reducción de escala para ver el comportamiento de las parametrizaciones con más de 60 parámetros.*

por lo que los parámetros de calidad de voz extraídos automáticamente sufren de una acusada falta de estabilidad, dando lugar a una parametrización poco robusta y con gran dispersión intra-emoción.

En general, y según estos resultados de desarrollo, se puede deducir que la mejor opción es combinar todas las parametrizaciones supra-segmentales, lo que proporciona un 58,2 % de precisión con 155 parámetros. Los estadísticos de espectro y la prosodia por separado alcanzan valores algo inferiores, aunque comparables entre sí (56,1 % y 55,7 % respectivamente).

Respecto a las parametrizaciones segmentales (Figura 5.3), puede comprobarse que tampoco alcanzan una saturación real, aunque al igual que ocurría con la base de datos *Berlin*, el incremento de la precisión es muy pequeño a partir de 20 parámetros. El comportamiento general de estas parametrizaciones es muy si-

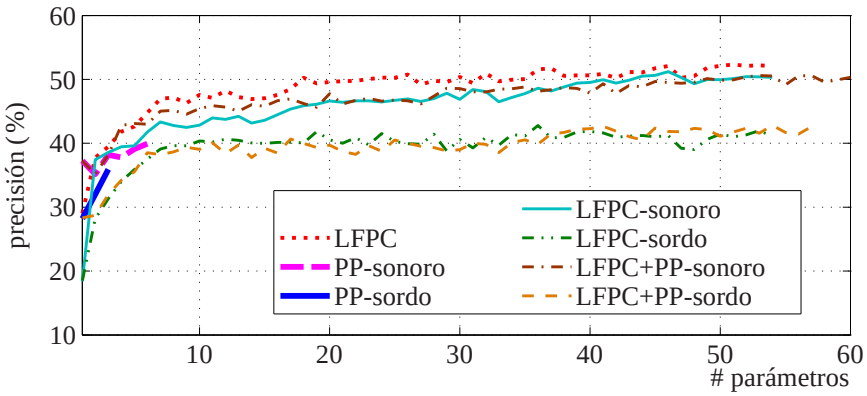


FIGURA 5.3: *Resultados de las pruebas de desarrollo para las parametrizaciones segmentales en la base de datos AIBO. El gráfico muestra la precisión del sistema en función del número de parámetros.*

milar al observado en el caso de las emociones actuadas, con los LFPC alcanzando la máxima precisión (59,6% con 20 parámetros), seguidos de cerca por los flujos sonoros de LFPC y de la combinación de LFPC y primitivas de prosodia (56,6% y 57,7% respectivamente). Los respectivos flujos sordos se encuentran bastante por debajo (50,5% y 49,7%).

Por el contrario, tomadas de forma aislada, las primitivas de prosodia consiguen resultados bastante pobres. Si bien en el caso de las emociones actuadas se observó que, para el mismo número de parámetros, la precisión obtenida por estas primitivas de prosodia era comparable a la de los LFPC, en este caso quedan muy por debajo. Tan sólo alcanzan un 49,9% y un 45,9% de precisión en los flujos sonoro y sordo respectivamente. Añadir estas primitivas de prosodia a la parametrización LFPC tampoco parece proporcionar ninguna ventaja.

Otra vez, resulta curioso que las primitivas de prosodia han sido, con diferencia, los parámetros mejor situados en el ranking de las parametrizaciones segmentales, al menos, para el flujo de tramas sonoras (Tabla 5.5). El problema, una vez más, radica en que no se debe deducir el comportamiento de los parámetros en conjunto a través de la evaluación de cada parámetro por separado. Las primitivas de prosodia son mucho más informativas que los LFPC cuando se toman de forma aislada, pero en conjunto, incluso para el mismo número de parámetros, los LFPC tienen mayor capacidad de discriminación. Si se observa el comportamiento de los sistemas con un único parámetro, puede comprobarse que, efectivamente, ese primer parámetro es más discriminante en el caso de las primitivas de prosodia (47,3%) que en el de los LFPC (28,4%). Sin embargo, la precisión obtenida con las primitivas prácticamente no crece, mientras que con los LFPC crece rápidamente.

5.4.2. Resultados independientes de locutor

La Tabla 5.6 muestra los resultados de la identificación automática de emociones en las pruebas independientes de locutor utilizando la base de datos *AIBO*. También muestra los resultados de las pruebas de desarrollo. Todos estos resultados se dan tanto para el número de parámetros óptimo estimado durante las pruebas de desarrollo como para el conjunto completo de parámetros.

En vista de estos resultados finales se puede comprobar, tal y como ya se había deducido de los resultados de desarrollo, que el añadir las características de calidad de voz a los parámetros prosódicos (columna 4) no aporta mejoras en la tasa de identificación con respecto a utilizar sólo la prosodia (columna 1). En ambos casos se consigue un 53,9 % de aciertos al usar los parámetros seleccionados y alrededor de un 54,5 % si se utilizan todos. Los estadísticos de espectro (columna 3) resultan ser la mejor parametrización suprasegmental aislada, alcanzando un 56 % de aciertos. Cuando se combinan todos los parámetros suprasegmentales (columna 5), la precisión no mejora significativamente.

Respecto a los parámetros segmentales, se comprueba que el flujo sordo de LFPC (columna 8) no es muy significativo a la hora de identificar las emociones naturales, ya que utilizando tan sólo el flujo sonoro (columna 7) se ha obtenido la misma precisión que con los LFPC sin separar los flujos (columna 6). También se aprecia que la incorporación de las primitivas de prosodia no proporciona ninguna ventaja con respecto a utilizar los LFPC aislados (columnas 11 y 12).

La Tabla 5.7 resume los resultados de la fusión tardía de expertos para la base de datos *AIBO*. Los sistemas en los que se hace uso de parámetros LFPC obtienen los mejores resultados, alrededor del 60 % de precisión cuando se utilizan todos los parámetros y entre el 58 % y el 60 % cuando se utilizan parámetros seleccionados. Es destacable que la tasa de aciertos conseguida con sólo el flujo sonoro de LFPC es precisamente de este mismo orden (Tabla 5.6, columna 7), lo que sugiere que el resto de parametrizaciones no ha aportado información añadida significativa. Entre los sistemas de fusión tardía que no utilizan parámetros LFPC sólo la combinación de la prosodia suprasegmental y las primitivas de prosodia se acerca a estos valores de precisión (columna 8). Sin embargo, la combinación de estos parámetros prosódicos con características espectrales a corto y largo plazo (columna 9) no proporciona ninguna mejora.

TABLA 5.6: *Precisión de las pruebas de desarrollo y finales sobre la base de datos* AIBO. *Se muestran los resultados con todos los parámetros y con el número óptimo de parámetros seleccionado durante el desarrollo. Valores de precisión en porcentaje.*

	Prosodia	**VQ**	**Espectral**	**Prosodia+ VQ**	**Prosodia+ VQ+ Espectral**	**LFPC**	**LFPC sonora**	**LFPC sorda**	**PP sonora**	**PP sorda**	**LFPC+ PP sonora**	**LFPC+ PP sorda**
# param.	25	4	119	11	155	20	20	20	6	3	20	20
Desarrollo	55,7	41,6	56,1	55,8	58,2	59,6	56,6	50,5	49,9	45,9	57,7	49.7
Final	53,9	38,3	56,6	53,9	56,7	56,1	57,3	47,4	49,0	44,8	57,1	47,6
# param.	54	5	324	59	383	54	54	54	6	3	60	57
Desarrollo	54,6	41,3	56,5	54,9	57,5	62,1	60,3	51,1	49,9	45,9	60,3	52.6
Final	54,6	38,3	56,0	54,4	56,0	60,1	60,1	50,2	49,0	44,8	59,2	51.7

TABLA 5.7: *Precisión de las pruebas de fusión tardía sobre la base de datos* AIBO. *Se muestran los resultados utilizando todos los parámetros y en el caso de utilizar el número óptimo de parámetros seleccionado durante el desarrollo. Valores de precisión en porcentaje.*

	Prosodia+ VQ+ Espectral	**LFPC-V+ LFPC-UV+ PP-V+PP-UV**	**LFPC-V+ LFPC-UV**	**PP-V+ PP-UV**	**Prosodia+ LFPC-V+ LFPC-UV**	**Espectral+ PP-V+ PP-UV**	**Espectral+ LFPC-V+ LFPC-UV**	**Prosodia+ PP-V+ PP-UV**	**TODOS**
Param. sel.	55,4	58,6	58,3	52,2	59,4	56,2	57,3	59,2	58,7
Todos param.	55,7	60,0	59,7	52,2	60,4	57,6	61,0	60,4	60,0

5.5. Pruebas con optimización cruzada

En la el capítulo 4 se han analizado los resultados de identificación automática de emociones actuadas utilizando la base de datos *Berlin*, mientras que en la sección 5.4 se han realizado experimentos similares para emociones naturales con la base de datos *AIBO*. En ambos casos los sistemas de identificación han sido optimizados mediante una serie de pruebas de desarrollo llevadas a cabo sobre emociones de la misma naturaleza que las pruebas finales. Es decir, los sistemas entrenados para emociones actuadas se han optimizado con resultados de emociones actuadas, y los entrenados para emociones naturales, con resultados de emociones naturales.

Entre las optimizaciones realizadas a los sistemas se incluyen el ranking de los parámetros y la selección del número óptimo de características a utilizar. Es interesante comprobar cuáles son los resultados obtenidos en la identificación de emociones naturales cuando la optimización del sistema se realiza con una base de datos de emociones actuadas y vice-versa. Esto permite determinar si las diferencias entre ambas optimizaciones son significativas o no, es decir, si un sistema diseñado para emociones de una cierta naturaleza puede ser utilizado con éxito sobre emociones de otra naturaleza. Concretamente se pretende comprobar si los parámetros seleccionados como óptimos son dependientes de la naturaleza de las emociones.

En las secciones 4.4 y 5.3 ya se ha comprobado que el ranking de parámetros es ligeramente diferente en ambos casos, y sería conveniente determinar hasta qué punto estas diferencias son realmente importantes. Para ello se han realizado pruebas de identificación automática de emociones sobre la base de datos *AIBO* utilizando el ranking y el número óptimo de parámetros establecido para *Berlin*, y pruebas de identificación automática de emociones sobre la base de datos *Berlin* utilizando el ranking y el número óptimo de parámetros establecido para *AIBO*.

No se trata de pruebas de bases de datos cruzadas, es decir, los clasificadores utilizados para los experimentos sobre la base de datos *Berlin* siguen estando entrenados mediante las señales de entrenamiento de esta base de datos, y lo mismo ocurre con los experimentos sobre *AIBO*. Sólo el ranking y número óptimo de parámetros se ha cruzado de una base de datos a la otra. Realizar pruebas cruzadas entre estas dos bases de datos no es viable, debido a que el conjunto de emociones que abarcan es diferente. Por ejemplo, no sería posible detectar voz *enfática* con un sistema entrenado con la base de datos *Berlin*, puesto que no recoge esta emoción. Sin embargo, las pruebas realizadas permiten estimar el efecto de la optimización de los parámetros y la importancia de sus diferencias. Si los resultados de precisión obtenidos mediante estas pruebas de optimización cruzada no tienen diferencias significativas con los descritos en las secciones anteriores, se puede deducir que las diferencias en la selección de parámetros en ambos casos no es

significativa. Esto, además, reforzaría la idea de que las conclusiones obtenidas para las emociones actuadas son válidas también para emociones naturales.

En este sentido es importante definir qué es una diferencia significativa. Para ello, se ha realizado una prueba χ^2 (Freund y Wilson, 1996) a cada uno de los resultados obtenidos, comparando los casos en los que la optimización se realiza sobre la misma base de datos y sobre la base de datos cruzada. Aquellos casos para los que la prueba proporciona un valor de $p < 0{,}05$ se consideran estadísticamente significativos.

Las Tablas 5.8 y 5.9 presentan las tasas de acierto logradas sobre la base de datos *Berlin* cuando los sistemas están optimizados para la base de datos *AIBO*. Las Tablas 5.10 y 5.11 recogen los resultados sobre la base de datos *AIBO* con los sistemas optimizados para *Berlin*. Como referencia también se proporcionan, en cada caso, los resultados obtenidos al optimizar los sistemas con la misma base de datos de pruebas, y que ya han sido comentados en secciones anteriores de este documento. Se puede comprobar que en la mayoría de las parametrizaciones, las diferencias en los resultados no son estadísticamente significativas. Como era de esperar, en aquellos casos en los que sí lo son el sistema optimizado para la base de datos objetivo logra mejores resultados.

También es apreciable que el número óptimo de parámetros estimado para cada una de las bases de datos es muy similar. Esto, junto con el hecho de que en la mayoría de los casos no haya diferencias significativas en los resultados, sugiere que, aunque el ranking de parámetros obtenido en ambas bases de datos es diferente, unas características son sustituidas por otras que proporcionan una información emocional similar. Por lo tanto, ambas listas son equivalentes en cuanto a su utilidad para la identificación de emociones. Este hecho, junto con los resultados experimentales presentados en la sección 5.4, refuerza la idea de que los análisis de parámetros llevados a cabo con ambas bases de datos son equivalentes, y que, por tanto, las conclusiones obtenidas para las emociones actuadas también son aplicables a las emociones naturales.

Hay que hacer notar que, aunque la precisión obtenida en la identificación de emociones naturales es muy inferior a la conseguida en emociones actuadas, estas diferencias se deben a la calidad de las señales (habla espontánea en un caso y grabaciones profesionales en el otro), y a que las emociones naturales no suelen darse con tanta intensidad como las actuadas. Estas diferencias no influyen en el hecho de que los parámetros que resultan ser buenos para una de las bases de datos también lo son para la otra.

TABLA 5.8: *Resultados de identificación sobre* Berlin *cuando se utilizan los parámetros optimizados para* AIBO. *Como referencia se dan los resultados obtenidos con la optimización sobre* Berlin. *Valores de precisión en porcentaje.*

Optimiz	Prosodia	VQ	Espectral	Prosodia+ VQ	Prosodia+ VQ+ Espectral	LFPC	LFPC sonora	LFPC sorda	PP sonora	PP sorda	LFPC+ PP sonora	LFPC+ PP sorda
AIBO	57,0	40,9	68,0	52,5	69,5	68,2	67,5	61,7	60,8	48,4	66,9	62,2
# param.	25	4	119	11	155	20	20	20	6	3	20	20
Berlin	62,6	47,3	70,5	60,9	72,2	69,9	63,2	61,5	60,8	48,4	66,4	57,0
# param.	39	5	96	17	152	20	20	20	6	3	20	20
Signif.†	–	∗	–	∗	–	–	–	–	–	–	–	–

† Relevancia estadística de las diferencias con $p = 0{,}05$ (∗) y $p = 0{,}01$ (∗∗).

TABLA 5.9: *Precisión de las pruebas de fusión tardía sobre* Berlin *cuando se utilizan los parámetros optimizados para* AIBO. *Como referencia se dan los resultados obtenidos con la optimización sobre* Berlin. *Valores de precisión en porcentaje.*

Optimiz	Prosodia+ VQ+ Espectral	LFPC-V+ LFPC-UV+ PP-V+PP-UV	LFPC-V+ LFPC-UV	PP-V+ PP-UV	Prosodia+ LFPC-V+ LFPC-UV	Espectral+ PP-V+ PP-UV	Espectral+ LFPC-V+ LFPC-UV	Prosodia+ PP-V+ PP-UV	TODOS
AIBO	66,2	76,5	73,3	65,4	72,5	73,1	72,9	68,4	76,1
Berlin	71,8	74,4	72,0	65,4	75,5	74,8	76,6	69,2	76,8
Signif.†	∗	–	–	–	–	–	–	–	–

† Relevancia estadística de las diferencias con $p = 0{,}05$ (∗) y $p = 0{,}01$ (∗∗).

TABLA 5.10: *Resultados de identificación sobre* AIBO *cuando se utilizan los parámetros optimizados para* Berlin. *Como referencia se dan los resultados obtenidos con la optimización sobre* AIBO. *Valores de precisión en porcentaje.*

Optimiz.	Prosodia	VQ	Espectral	Prosodia+ VQ	Prosodia+ VQ+ Espectral	LFPC	LFPC sonora	LFPC sorda	PP sonora	PP sorda	LFPC+ PP sonora	LFPC+ PP sorda
Berlin	54,7	38,3	56,7	54,6	58,4	56,0	56,4	43,1	49,0	44,8	57,1	45,5
# param.	39	5	96	17	152	20	20	20	6	3	20	20
AIBO	53,9	38,3	56,6	53,9	56,7	56,1	57,3	47,4	49,0	44,8	57,1	47,6
# param.	25	4	119	11	155	20	20	20	6	3	20	20
Signif.†	–	–	–	–	–	–	–	**	–	–	–	*

† Relevancia estadística de las diferencias con $p = 0{,}05$ ($*$) y $p = 0{,}01$ ($**$).

TABLA 5.11: *Precisión de las pruebas de fusión tardía sobre* AIBO *cuando se utilizan los parámetros optimizados para* Berlin. *Como referencia se dan los resultados obtenidos con la optimización sobre* AIBO. *Valores de precisión en porcentaje.*

Optimiz.	Prosodia+ VQ+ Espectral	LFPC-V+ LFPC-UV+ PP-V+PP-UV	LFPC-V+ LFPC-UV	PP-V+ PP-UV	Prosodia+ LFPC-V+ LFPC-UV	Espectral+ PP-V+ PP-UV	Espectral+ LFPC-V+ LFPC-UV	Prosodia+ PP-V+ PP-UV	TODOS
Berlin	55,3	58,5	56,3	52,2	57,0	56,8	58,7	54,4	59,2
AIBO	55,4	58,6	58,3	52,2	59,4	56,2	57,3	59,2	58,7
Signif.†	–	–	*	–	*	–	–	**	–

† Relevancia estadística de las diferencias con $p = 0{,}05$ ($*$) y $p = 0{,}01$ ($**$).

5.6. Conclusiones

En este capítulo se ha utilizado una base de datos de emociones naturales y habla espontánea para verificar si las conclusiones extraídas en el capítulo 4 son generalizables a situaciones más realistas.

Los resultados obtenidos, tanto con el análisis de variabilidad como con las pruebas experimentales, sugieren que los parámetros derivados de la calidad de voz no son adecuados para este tipo de señales de voz espontánea. La razón no es que no contengan información acerca de la emoción del locutor, sino que no se ha encontrado la manera de parametrizar esta calidad de voz de forma robusta. La extracción automática de características de calidad de voz es compleja, incluso en señales limpias grabadas bajo condiciones controladas, y requiere muchas veces de supervisión humana. El aplicar estas técnicas a voces espontáneas proporciona valores poco robustos que no hacen sino confundir al clasificador. Mientras que en la base de datos de emociones actuadas los parámetros de calidad de voz podían proporcionar información añadida si se utilizaban junto con características prosódicas, en el caso de las emociones naturales espontáneas no aportan ninguna mejora.

Tampoco los valores de primitivas de prosodia parecen adecuados para este tipo de señales. Al igual que ocurre con la calidad de voz, existen problemas a la hora de aplicar algoritmos de estimación de curvas de entonación sobre señales de voz espontánea, debido a que generalmente no tienen una pronunciación cuidada. Se ha comprobado que el combinar las primitivas de prosodia y las parametrizaciones LFPC no mejora los resultados, ni mediante fusión temprana ni fusión tardía.

Aunque los valores de primitivas de prosodia no parezcan adecuados, los parámetros prosódicos supra-segmentales tradicionales parecen capturar de forma más adecuada la información emocional. Debido a que estos parámetros están calculados mediante una ventana de integración larga, es de esperar que el efecto de los errores cometidos durante el proceso de estimación de las curvas primitivas se suavice, proporcionando una parametrización más robusta. Sin embargo, la precisión obtenida sigue siendo inferior a la conseguida con los estadísticos a largo plazo de la envolvente espectral, tal y como ocurría en el caso de la base de datos actuada.

Mientras que en las pruebas sobre la base de datos *Berlin* utilizar la fusión tardía para combinar información de una misma fuente pero con intervalos de integración diferentes (a nivel de trama y a nivel de frase) aumentaba la precisión con respecto a utilizar una única base de tiempos, en este caso sólo se ha encontrado evidencias de este comportamiento con las características asociadas a la prosodia. La combinación de estadísticos prosódicos a largo plazo y las primitivas de prosodia logra una precisión que se sitúa entre las más elevadas. Por el contrario,

no parece que la combinación de parámetros espectrales de diferente naturaleza temporal proporcione ventajas con respecto a utilizar los LFPC de forma aislada.

A grandes rasgos, las conclusiones principales obtenidas durante este análisis realizado sobre emociones naturales se resumen en la siguiente lista:

- Las características extraídas de los estadísticos de prosodia son más significativas para la identificación de emociones si se consideran individualmente o en grupos reducidos.
- En caso de considerar un mayor número de parámetros, las características espectrales proporcionan una mayor separación de las emociones.
- Los parámetros de calidad de voz calculados de forma automática no son adecuados para el habla espontánea, al menos, los considerados en este trabajo. Los errores cometidos durante la parametrización provocan grandes dispersiones en las características, aumentando la confusión del clasificador. En caso de querer utilizar características de calidad de voz, será necesario realizar una revisión manual de los parámetros o mejorar los sistemas de caracterización.
- Al contrario de lo que ocurre con las emociones actuadas, la combinación de las características supra-segmentales no proporciona ninguna mejora significativa con respecto a utilizar tan sólo el espectro.
- Utilizar la fusión tardía para combinar información extraída de una misma fuente pero con intervalos de integración diferentes (a nivel de trama y a nivel de frase) sólo permite mejorar la precisión con los parámetros de intensidad y entonación. La mejora con parámetros prosódicos no es significativa.
- La parametrización LFPC consigue por sí misma los mejores resultados[2]. El flujo sonoro de LFPC consigue prácticamente los mismos resultados.

La mayoría de estas conclusiones coinciden con las obtenidas durante el análisis de las emociones actuadas. La mayor diferencia está en que, en las emociones naturales, la importancia de las parametrizaciones espectrales es aún más acusada que en las emociones actuadas. En las parametrizaciones supra-sementales, la combinación de prosodia y estadísticos de envolvente espectral no mejora nada con respecto a utilizar sólo las características espectrales. Y en las segmentales, el uso de LFPC ya consigue la máxima precisión posible. Los parámetros derivados de la prosodia sólo son útiles si se toman muy pocos parámetros.

[2] Aunque hay algunos resultados de fusión tardía que son ligeramente mejores, las diferencias no son significativas.

Esto confirma una vez más que, aunque los parámetros prosódicos presentan una gran capacidad para discriminar emociones si se toman individualmente, no son los mejores cuando se considera todo el conjunto de parámetros. En el caso concreto de las emociones naturales y habla espontánea, las características prosódicas tienen además el inconveniente de estar estimadas de forma menos robusta, debido a la falta de cuidado en la pronunciación y vocalización. En el caso de las emociones actuadas, la prosodia gana cierto protagonismo gracias a que las curvas de entonación pueden ser calculadas con mayor precisión.

También se han realizado experimentos de identificación de emociones con optimización cruzada. Se ha comprobado que, aunque las dos bases de datos proporcionan un ranking de parámetros diferente, los resultados finales son muy similares a los obtenidos con los sistemas optimizados para la base de datos objetivo, con muy pocas diferencias significativas. Esto induce a pensar que, efectivamente, el ranking resultante en ambas bases de datos es equivalente en cuanto a su aplicación para la identificación de emociones.

Éste último resultado es muy interesante, ya que permite utilizar una base de datos de emociones actuadas para diseñar un sistema de clasificación de emociones naturales. Las bases de datos de emociones actuadas son más sencillas de obtener, y permiten controlar el contenido emocional de las grabaciones, eliminando la necesidad de realizar un etiquetado. Una vez conseguidas las grabaciones con el contenido emocional requerido, la selección de los parámetros a utilizar puede realizarse sobre esta base de datos actuada, bien por análisis de las características de cada parametrización o bien mediante pruebas experimentales de identificación automática. Según los resultados obtenidos en esta tesis, los parámetros optimizados para estas emociones actuadas proporcionan resultados equivalentes a los obtenidos si se optimizaran sobre las emociones naturales en las que se va a utilizar el sistema.

Capítulo 6

Conclusiones

Índice

ESTA tesis presenta un análisis sistemático de las parametrizaciones más comúnmente utilizadas en la identificación automática de emociones, con el objetivo de determinar su efectividad en esta tarea. En la literatura existen varios trabajos que analizan la capacidad proporcionada por diferentes parámetros para la discriminación de emociones. Sin embargo, no presentan una visión completa del comportamiento de estas parametrizaciones. Muchos de estos trabajos analizan cada parámetro de forma individual, lo que puede dar lugar a interpretaciones no válidas cuando estos parámetros se utilizan en combinación con otros. Otros estudios proporcionan tasas de precisión obtenidas mediante evaluación experimental para una cierta parametrización, con lo que se puede deducir el comportamiento global de todo el conjunto de parámetros. Sin embargo, no se suelen proporcionar los resultados de las parametrizaciones por separado, dando sólo los resultados de la combinación total, con lo que no es posible saber si esta combinación realmente es mejor que las partes por separado. Además, los diferentes estudios presentes en la literatura no son comparables entre sí, debido a las diferencias en la arquitectura de los experimentos, el número y tipo de emociones, la calidad de las grabaciones u otros factores. Por lo tanto, no es posible construir una visión completa de las propiedades de cada parametrización simplemente comparando los resultados de los diferentes trabajos publicados.

Esta falta de un análisis exhaustivo de las parametrizaciones provoca que la comunidad investigadora no llegue a un acuerdo sobre cuáles son los parámetros más relevantes para la identificación de emociones en el habla. Con el estudio presentado en esta tesis se ha tratado de llenar este vacío, mostrando resultados comparables entre diferentes parámetros, y su utilidad para la identificación de emociones.

6.1. Aportaciones de la tesis y trabajos futuros

6.1.1. Análisis de parámetros

Se ha realizado un análisis sistemático de parámetros de naturaleza espectral, prosódica y de calidad de voz con respecto a su utilidad para la identificación automática de emociones en la voz. Este análisis se ha llevado a cabo mediante diferentes métodos, estudiando cada parámetro por separado (a través de técnicas de ranking de parámetros), así como considerando todo el conjunto de parámetros (mediante medidas de dispersión multidimensional). También se han realizado pruebas experimentales de identificación automática, con el objetivo de validar las conclusiones obtenidas. Además, todos estos análisis se han llevado a cabo utilizando tanto emociones actuadas como naturales, llegando en ambos casos a conclusiones similares.

Los resultados desvelan que, en contra de la corriente más aceptada dentro de la comunidad científica, los parámetros prosódicos o de calidad de voz más habitualmente utilizados no son los más adecuados para esta tarea, ya que las características espectrales presentan mayor capacidad de discriminación. Tampoco la combinación de características espectrales, prosódicas y de calidad de voz proporciona una mejora apreciable de las tasas de precisión con respecto a utilizar sólo características espectrales.

Probablemente el escaso rendimiento de los parámetros prosódicos y de calidad de voz es debida a la dificultad en el cálculo de estos parámetros a partir de la señal de voz, sobre todo, en voz espontánea. Aunque se considera que las características prosódicas y de calidad de voz son un vehículo importante de la información emocional, no es sencillo extraer esta información mediante algoritmos automáticos. Esto hace que los parámetros estimados sean poco robustos y confundan al sistema de clasificación. Por el contrario, las características espectrales muestran mayor estabilidad, a la vez que transportan una cantidad de información emocional considerable. Esto hace que al final presenten mejores cualidades para la identificación.

Como trabajo futuro se plantea incrementar este análisis a otro tipo de parámetros, no considerados durante este estudio. Concretamente, sería interesante determinar la aportación de parámetros lingüísticos en los sistemas de identificación de emociones, y comparar su comportamiento con el resto de parametrizaciones acústicas. Los parámetros lingüísticos pueden ser muy adecuados en sistemas que trabajan con emociones espontáneas, y pueden ayudar a mejorar significativamente la tasa de identificación, que en este tipo de emociones suele ser bastante baja.

6.1.2. Comparación de metodologías de fusión de información

Se ha comparado el resultado de la fusión temprana (concatenación de parámetros) y la fusión tardía (mezcla de scores) para la combinación de información obtenida a partir de parámetros de diferente naturaleza acústica. Los resultados obtenidos reflejan que, en términos de precisión final, ambos mecanismos son equivalentes. Sin embargo, la fusión temprana sólo puede aplicarse a aquellas parametrizaciones que comparten una misma estructura temporal, es decir, que por cada vector calculado por una de las parametrizaciones exista otro vector en la otra, calculado para el mismo instante de tiempo. Por el contrario, la fusión tardía permite combinar la información de parametrizaciones con estructuras temporales muy diferentes. Por ejemplo, permite fusionar características segmentales con supra-segmentales, o parámetros de tramas sonoras con parámetros de tramas sordas.

A partir de los resultados se deduce que la combinación de parámetros de la misma naturaleza acústica (prosódica o espectral) pero de diferente estructura temporal es beneficiosa para la correcta identificación de las emociones. Puesto que este tipo de combinaciones sólo puede realizarse mediante fusión tardía, y en los casos en los que la fusión temprana es viable, los resultados son equivalentes, puede concluirse que la combinación por mezcla de scores es más adecuada para esta tarea.

Se propone analizar otras técnicas de fusión no consideradas en este estudio, como por ejemplo, la combinación jerárquica de clasificadores (Ruta y Gabrys, 2000). Para los casos en los que se quieran fusionar muchas fuentes de información diferentes (en este estudio hay un caso en el que se combinan 7 parametrizaciones), es posible realizar la fusión de forma jerárquica en lugar de combinar todas las parametrizaciones en una sola etapa. Por ejemplo, se pueden combinar las fuentes de información por parejas, luego estos resultados también por parejas, así hasta acabar con un único resultado final. La ventaja de este método es que es posible que cada sistema de fusión aprenda a aprovechar con mayor detalle la información agregada de cada pareja de parametrizaciones, dando como resultado una menor tasa de error. La desventaja es que la complejidad del sistema aumenta en gran medida, y que es necesario disponer de muchos más datos de entrenamiento para poder diseñar la fusión.

6.1.3. Algoritmos de extracción de características

En todo momento se han querido evaluar parametrizaciones obtenidas mediante algoritmos totalmente automáticos, sin supervisión humana. La razón fundamental es que, en una aplicación final destinada a la identificación automática de emociones, generalmente no es posible realizar ajustes manuales en los pará-

metros, por lo que todo el sistema depende de algoritmos automáticos. Para poder llevar a cabo esta parametrización con cierto grado de fiabilidad, ha sido necesario diseñar nuevos algoritmos para el procesado de la señal de voz, así como modificar otros ya existentes.

Se ha diseñado un nuevo sistema para la estimación de las curvas de entonación basado en análisis cepstrum y programación dinámica (CDP), que ha demostrado ser especialmente robusto tanto en señales libres de ruido como en señales de baja SNR. También se ha implementado un sistema de marcado a período de pitch utilizando como referencia la curva de entonación estimada mediante este algoritmo y la curva glotal estimada mediante filtrado inverso de la señal de voz.

Además, se ha diseñado un sencillo mecanismo para la detección automática de vocales basado en un sistema de reconocimiento utilizando modelos acústicos de agrupaciones fonéticas. La agrupación fonética es un elemento esencial en este sistema, ya que permite modelar conjuntamente grupos de fonemas de características acústicas similares, reduciendo la confusión del sistema. Gracias a este sistema se ha conseguido detectar automáticamente en la señal la posición de segmentos de naturaleza vocálica, lo que permite realizar una rápida estimación de la velocidad del habla, una característica que puede variar de una emoción a otra. Los segmentos detectados también han sido utilizados para el cálculo de características asociadas a la calidad de voz, pues las vocales son regiones de gran estabilidad, lo cual es fundamental para poder realizar una estimación robusta de la señal glotal.

Por último, se ha mejorado uno de los sistemas de detección de actividad vocal más conocidos, basado en LTSE, con el objetivo de adaptarlo mejor a condiciones de ruido ambiente. Las modificaciones realizadas sobre este algoritmo permiten reducir el número de tramas de silencio que son clasificadas como voz a la vez que mantienen el número de tramas de voz clasificadas como silencio dentro de unos márgenes aceptables para aplicaciones de identificación de emociones. Estos cambios también resultarían beneficiosos en aquellas aplicaciones en las que los segmentos de silencio provocan un incremento de la tasa de errores, como pueden ser la identificación de locutores o de idioma.

En el futuro sería muy interesante perfeccionar los algoritmos de estimación de parámetros de calidad de voz, ya que se ha comprobado que proporcionan valores poco robustos, sobre todo en voz espontánea. Obtener una parametrización más fiable puede hacer que la información emocional contenida sea útil en los sistemas de identificación automática de emociones. Este perfeccionamiento puede conseguirse a través de la mejora de las técnicas de filtrado inverso, o mediante la definición de nuevos parámetros que puedan calcularse de forma más robusta en señales de poca calidad.

6.2. Difusión de resultados

Artículos de revista

Navas, E., Hernáez, I., and Luengo, I. (2006). An objective and subjective study of the role of semantics in building corpora for TTS. *IEEE transactions on Speech and Audio Processing*, 14(4):1117–27.

Luengo, I., Navas, E., and Hernáez, I. Feature analysis and evaluation for automatic emotion identification in speech. Aceptado para publicación en *IEEE Transactions on Multimedia*.

Luengo, I., Navas, E., Sánchez, J., and Hernáez, I. (2009). Detección de vocales mediante modelado de clusters de fonemas. *Procesado del Lenguaje Natural*, 43:121–128.

Navas, E., Hernáez, I., Luengo, I., Sainz, I., Saratxaga, I., and Sanchez, J. (2007). Meaningful parameters in emotion characterisation. *Lecture Notes on Artificial Intelligence*, 4775:74–84.

Navas, E., Hernáez, I., Luengo, I., Sánchez, J., and Saratxaga, I. (2005). Analysis of the suitability of common corpora for emotional speech modeling in standard Basque. *Lecture Notes on Artificial Intelligence*, 3658:265–272.

Navas, E., Hernáez, I., Castelruiz, A., and Luengo, I. (2004). Obtaining and evaluating an emotional database for prosody modelling in standard Basque. *Lecture Notes on Artificial Intelligence*, 3206:393–400.

Navas, E., Hernáez, I., Castelruiz, A., Sánchez, J., and Luengo, I. (2004). Acoustic analysis of emotional speech in standard Basque for emotion recognition. *Lecture Notes on Computer Science*, 3287:386–393.

Participación en congresos

Luengo, I., Navas, E., Odriozola, I., Saratxaga, I., Hernáez, I., Sainz, I., Error, D. Modified LTSE VAD algorithm for applications requiring reduced silence frame misclassification. Aceptado para *Language Resources and Evaluation Conference (LREC)*, 2010

Luengo, I., Navas, E., and Hernáez, I. (2009). Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. En *Interspeech*, páginas 332–335, Brighton, Reino Unido.

Sainz, I., Saratxaga, I., Navas, E., Hernáez, I., Sánchez, J., Luengo, I., Odriozola, I., and de Bilbao, E. (2008). Evaluación subjetiva de una base de datos de habla emocional para euskera. En *Jornadas de Tecnología del Habla (JTH)*, páginas 191–194, Bilbao, España.

Sainz, I., Saratxaga, I., Navas, E., Hernáez, I., Sánchez, J., Luengo, I., and Odriozola, I. (2008). Subjective evaluation of an emotional speech database for Basque. En *Language Resources and Evaluation Conference (LREC)*, páginas 1712–1715, Marrakech, Marruecos.

Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sánchez, J., and Sainz, I. (2007). Evaluation of pitch detection algorithms under real conditions. En *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, páginas 1057–1060, Honolulu, EEUU.

Saratxaga, I., Navas, E., Hernáez, I., and Luengo, I. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. En *Language Resources and Evaluation Conference (LREC)*, páginas 2126–2129, Génova, Italia.

Luengo, I., Navas, E., Hernáez, I., and Sanchez, J. (2005). Automatic emotion recognition using prosodic parameters. En *Interspeech*, páginas 493–496, Lisboa, Portugal.

Bibliografía

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723.

Alkoot, F. M. and Kittler, J. (1999). Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20:1361–69.

Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118.

Amir, N., Ron, S., and Laor, N. (2000). Analysis of an emotional speech corpus in hebrew based on objective criteria. In *ISCA workshop on speech and emotion*, pages 29–33, Newcastle, UK.

Ayat, N. E., Cheriet, M., and Suen, C. Y. (2005). Automatic model selection for the optimization of SVM kernels. *Patter Recognition*, 38(10):1733–1745.

Bäckström, T., Alku, P., and Vilkman, E. (2002). Time–domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range. *IEEE transactions on Speech and Audio Processing*, 10:186–192.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Pathology*, 70(3):614–636.

Barra-Chicote, R., Fernandez, F., Lutfi, S., Lucas-Cuesta, J., Macias-Guarasa, J., Montero, J., San-Segundo, R., and Pardo, J. (2009). Acoustic emotion recognition using dynamic bayesian networks and multi-space distributions. In *Interspeech*, pages 336–339, Brighton, UK.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards and human beings. In *ISCA workshop on speech and emotion*, pages 195–200, Belfast, Irlanda del Norte.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2006). Combining

efforts for improving automatic classification of emotional user states. In *Information Society - Language TechnologiesConference (IS-LTC)*, pages 240–245, Ljubljana (Slovenia).

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *Journal on Applied Signal Processing*, 4:430–451.

Boersma, P. (1993). Accurate short term analysis of the fundamental frequency and the harmonics-to-noise ratio. In *Institute of Phonetic Sciences*, volume 17, pages 97–100, University of Amsterdam.

Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication*, 40:213–225.

Bozkurt, E., Erzin, E., Erdem, Ç. E., and Erdem, A. T. (2009). Improving automatic emotion recognition from speech signals. In *Interspeech*, pages 324–327, Brighton, UK.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of German emotional speech. In *Interspeech*, pages 1517–1520, Lisbon. Portugal.

Burkhardt, F. and Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop on Speech and Emotion*, pages 151–156, Belfast.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462.

Campbell, N. (2000). Databases of emotional speech. In *ISCA workshop on speech and emotion*, pages 34–38, Belfast, Irlanda del Norte.

Casale, S., Russo, A., and Serano, S. (2007). Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, 49(10):801–810.

Chang, C.-C. and Lin, C.-J. (2004). LIBSVM: A library for support vector machines.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine learning*, 46:131–159.

Chen, J., Wang, C., and Wang, R. (2009). Adaptive binary tree for fast SVM multiclass classification. *Neurocomputing*, 72:3370–3375. Hybrid Learning Machines (HAIS 2007) / Recent Developments in Natural Computation (ICNC 2007).

Chen, P.-H., Lin, C.-J., and Schölkopf, B. (2005). A tutorial on nu-support vector machines. *Applied Stochastic Models in Business and Industry*, 21:111–136.

Chichosz, J. and Slot, K. (2007). Emotion recognition in speech signal using emotion-extracting binary decision trees. In *Affective Computing and Intelligent Interfaces (ACII) 2007*, Lisbon, Portugal.

Choukri, K. (2003). *Groningen database*. European Language Resources Association (ELRA). Catalog reference S0020, www.elra.info.

Cornelius, R. R. (2000). Theoretical approaches to emotion. In *ISCA workshop on speech and emotion*, pages 3–10, Belfast, Irlanda del Norte.

Cowie, R. (2000). Describing the emotional states expressed in speech. In *ISCA workshop on speech and emotion*, pages 11–18, Belfast, Irlanda del Norte.

Cowie, R., Douglas-Cowie, E., and Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. New York.

Devillers, L., Vasilescu, I., and Vidrascu, L. (2004). F0 and pause features analysis for anger and fear detection in real-life spoken dialogs. In *Speech Prosody*, pages 205–208, Nara, Japan.

Dornaikaa, F. and Raducanub, B. (2007). Inferring facial expressions from videos: Tool and application. *Signal Processing: Image Communication*, 22(9):769–784.

Douglas-Cowie, E., Cowie, R., and Schröder, M. (2000). A new emotional database: Considerations, sources and scope. In *ISCA workshop on speech and emotion*, pages 39–44, Belfast, Irlanda del Norte.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6:169–200.

Engberg, I. S. and Hansen, A. V. (1996). Documentation of the Danish emotional speech database (DES). Internal aau report, Center for Person Kommunikation, Denmark.

Erickson, D. (2005). Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, 26(4):317–325.

ETSI (1997). *ES 301 249: Digital cellular telecommunications system (Phase 2); Voice Activity Detector (VAD) for Enhanced Full Rate (EFR) speech traffic channels (GSM 06.82 version 4.0.1)*.

ETSI (2003). *ES 202 050: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*.

Fernandez, R. and Picard, R. W. (2000). Modeling drivers' speech under stress. In *ISCA Tutorial and Research Workshop on Speech and Emotion*, pages 219–224, Newcastle, UK.

Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., and Gonzalez-Rodriguez, J. (2005). Adapted user-dependent multimodal biometric authentication exploiting general information. *Pattern Recognition Letters*, 26(16):2628–2639.

Freund, R. J. and Wilson, W. J. (1996). *Statistical Methods*. Academic Press, San Diego, USA.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Boston, USA.

Gobl, C. and Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40:189–212.

Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10):787–800.

Gutschoven, B. and Verlinde, P. (2000). Multi-modal identity verification using support vector machines (SVM). In *3rd Int. Conf. on Information Fusion*, volume 2, pages 3–8, Paris, France.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hansen, J. H. and Bou-Ghazale, S. (1997). Getting started with SUSAS: A speech under simulated and actual stress database. In *Eurospeech*, volume 4, pages 1743–1746, Rhodes, Greece.

Hashizawa, Y., Takeda, S., Muhd, D. H., and Ghen, O. (2004). On the differences in prosodic features of emotional expressions in Japanese speech according to the degree of the emotion. In *Speech Prosody*, pages 655–658, Nara, Japan.

Hernáez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., and Sánchez, J. (2003). The Basque Speech–Dat (II) database: A description and first test recognition results. In *Eurospeech*, pages 1549–1552, Geneva.

Hess, W. (1983). *Pitch determination of speech signals: algorithms and devices*. Springer, Berlin, Germany.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Hosom, J. P. (2000). *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology.

Hozjan, V. and Kacic, Z. (2003). Improved emotion recognition with large set of statistical features. In *Eurospeech*, pages 133–136, Ginebra, Suiza.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., and Nogueiras, A. (2002). Interface database: Design and collection of a multilingual emotional speech database. In *3rd Language Resources and Evaluation Conference*, pages 2024–2028, Las Palmas de Gran canaria, Spain.

Hsu, C. W. and Lin, C. J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.

Huang, C. F. and Akagi, M. (2008). A three-layered model for expressive speech perception. *Speech Communication*, 50(10):810–828.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., and Longhi, L. (2000). Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In *ISCA workshop on speech and emotion*, pages 161–166, Belfast, Irlanda del Norte.

Iskra, D., Grosskopf, B., Marasek, K., van del Heuvel, H., Diehl, F., and Kiessling, A. (2002). SPEECON – speech databases for consumer devices: database specification and validation. In *Language Resources and Evaluation Conference (LREC)*, pages 329—-333, Las Palmas, Spain.

ITU-T (2007). *Recommendation G.729 Annex B: A silence compression scheme for G.729 optimized for terminals conforming to ITU-T Recommendation V.70*.

Jiang, D.-N. and Cai, L.-H. (2004). Classifying emotion in Chinese speech by decomposing prosodic features. In *Interspeech*, pages 1325–1328, Jeju (Korea).

Johnstone, T. and Scherer, K. R. (1999). The effects of emotions on voice quality. In *International Conference of Phonetic Sciences*, pages 2029–2032, San Francisco, USA.

Keerthi, S. S. and Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computing*, 15(7):1667–1689.

Kienast, M. and Sendlmeier, W. F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA workshop on speech and emotion*, pages 92–97, Newcastle, UK.

Kim, S., Georgiou, P. G., Lee, S., and Narayanan, S. (2007). Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *IEEE Workshop on Multimedia Signal Processing*, pages 48–51, Crete.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–39.

Kwon, O. W., Chan, K., Hao, J., and Lee, T. W. (2003). Emotion recognition by speech signals. In *Eurospeech*, pages 125–128, Geneva, Switzerland.

Lee, C.-C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2009). Emotion recognition using a hierarchical binary decision tree approach. In *Interspeech*, pages 320–324, Brighton, UK.

Lee, C. M., Narayanan, S., and Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. In *Automatic Speech Recognition and Understanding Workshop*, pages 240– 243, Trento, Italia.

Liberman, M., Davis, K., Grossman, M., Martey, N., and Bell, J. (1999). *Emotional prosody speech and transcripts database*. Linguistic Data Consortium (LDC). Catalog reference LDC2002S28, http://www.ldc.upenn.edu/Catalog/.

Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recogniser based on SPEECHDAT(II). In *ICSLP*, volume 3, pages 370–373, Beijing.

López-Cozar, R., Callejas, Z., Kroul, M., Nouza, J., and Silovský, J. (2008). Two-level fusion to improve emotion classification in spoken dialogue systems. *Lecture Notes on Computer Science*, (5246):617–624.

Lu, X. and Dang, J. (2008). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Communication*, 50(4):312–322.

Luengo, I., Navas, E., and Hernáez, I. (2010). Modified LTSE VAD algorithm for applications requiring reduced silence frame misclassification. In *Language Resources and Evaluation Conference (LREC)*, page (To appear).

Luengo, I., Navas, E., and Hernáez, I. (2009a). Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. In *Interspeech*, pages 332–335, Brighton, UK.

Luengo, I., Navas, E., Hernáez, I., and Sanchez, J. (2005). Automatic emotion recognition using prosodic parameters. In *Interspeech*, pages 493–496, Lisbon, Portugal.

Luengo, I., Navas, E., Sánchez, J., and Hernáez, I. (2009b). Detección de vocales mediante modelado de clusters de fonemas. *Procesado del Lenguaje Natural*, 43:121–128.

Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sánchez, J., and Sainz, I. (2007). Evaluation of pitch detection algorithms under real conditions. In *ICASSP*, pages 1057–1060, Honolulu, USA.

Lugger, M. and Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICCASP'07)*, volume 4, pages 17–20, Honolulu, USA.

Madzarov, G., Gjorgjevikj, D., and Chorbev, I. (2009). A multi-class SVM classifier utilizing binary decision tree. *Informatica*, 33:233–241.

Makarova, V. and Petrushin, V. A. (2002). RUSLANA: A database of Russian emotional utterances. In *International Conference on Spoken Language Processing (ICSLP)*, pages 2041–2044, Denver, USA.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., and Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *ISCA workshop on speech and emotion*, pages 207–212, Belfast, Irlanda del Norte.

Medan, Y., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 39:40–48.

Montero, J., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., and Pardo, J. (1999). Analysis and modelling of emotional speech in Spanish. In *International Conference of Phonetic Sciences*, volume 2, pages 957–960, San Francisco, USA.

Morrison, D., Wang, R., and De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112.

Mozziconacci, S. J. and Hermes, D. J. (2000). Espression of emotion and attitude through temporal speech variations. In *International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 373–378, Beijing, China.

Müller, R., Schuller, B., and Rigoll, G. (2004). Enhanced robustness in speech emotion recognition combining acoustic and semantic analyses. In *From Signals To Signs of Emotion and Vice Versa*, Santorino, Greece.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel based learning algorithms. *Transactions on Neural Networks*, 12(2):181–202.

Navas, E., Hernáez, I., Castelruiz, A., Sánchez, J., and Luengo, I. (2004a). Acoustic analysis of emotional speech in standard Basque for emotion recognition. *Lecture Notes on Computer Science*, 3287:386–393.

Navas, E., Hernáez, I., Castelruiz, A., and Luengo, I. (2004b). Obtaining and evaluating an emotional database for prosody modelling in standard Basque. *Lecture Notes on Artificial Intelligence*, 3206:393–400.

Nicholson, J., Takahashi, K., and Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing and Applications*, 9(4):290–296.

Nogueiras, A., Moreno, A., Bonafonte, A., and Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. In *Eurospeech*, pages 2679–2682, Aalborg, Denmark.

Nwe, T. L., Foo, S. W., and de Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4):603–623.

Orman, O. D. and Arslan, L. M. (2001). Frequency analysis of speaker identification. In *Speaker Odyssey 2001*, pages 219–222.

Osuna, E. E., Freund, R., and Girosi, F. (1997). Support vector machines: Training and applications. Technical Report 1602, MIT.

Paalanen, P., Kamarainen, J.-K., Ilonen, J., and Kälviäinen, H. (2006). Feature representation and discrimination based on gaussian mixture model probability densities - practices and algorithms. *Pattern Recognition*, 39(7):1346–1358.

Paeschke, A. (2004). Global trend of fundamental frequency in emotional speech. In *Speech Prosody*, pages 671–674, Nara, Japan.

Paeschke, A., Kienast, M., and Sendlmeier, W. F. (1999). F0-contours in emotional speech. In *14th International Conference of Phonetic Sciences (ICPhS'99)*, page 929–932, San Francisco, USA.

Pellegrino, F. and Andre-Obrecht, R. (2000). Automatic language identification: an alternative approach to phonetic modelling. *Signal Processing*, 7.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Petrushin, V. A. (2000). Emotion recognition in speech signal: Experimental study, development and application. In *ICSLP*, pages 222–225, Beijing, China.

Pfau, T. and Ruske, G. (1998). Estimating the speaking rate by vowel detection. In *International Conference on Acoustics, Speech, and Signal Processing (ICCASP'98)*, pages 945–948.

Pierre-Yves, O. (2003). The production and recognition of emotions in speech: Features and algorithms. *Int. Journal of Human-Computer Studies*, 59:157–183.

Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., and Metze, F. (2009). Emotion classification in children's speech using fusion of acoustic and linguistic features. In *Interspeech*, pages 340–344.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–86.

Rabiner, L. and Schafer, R. (1978). *Digital processing of speech signals*. Signal Processing. Prentice Hall, New Jersey.

Ramirez, J., Segura, J. C., Benitez, C., de la Torre, A., and Rubio, A. (2004). Efficient voice activity detection algorithms using long term speech information. *Speech Communication*, 42:271–287.

Reeves, B. and Clifford, N. (2003). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*.

Reynolds, D. A. and Rose, R. C. (1995). Robust text independent speaker identification using gaussian mixture speaker models. *IEEE transactions on Speech and Audio Processing*, 3:72–83.

Riegelsberger, E. L. and Krishnamurthy, A. K. (199). Glottal source estimation: Methods of applying the LF-model to inverse filtering. In *International Conference on Acoustics, Speech, and Signal Processing (ICCASP'93)*, volume 2, pages 27–30.

Ringeval, F. and Chetouani, M. (2008). Exploiting a vowel based approach for acted emotion recognition. *Lecture Notes on Computer Science*, 5042:243–254.

Ruta, D. and Gabrys, B. (2000). An overview of classifier fusion methods. *Computing and Information Systems*, 7(1):1–10.

Sanchez A., V. D. (2003). Advanced support vector machines and kernel methods. *Neurocomputing*, 55:5–20.

Saratxaga, I., Navas, E., Hernáez, I., and Luengo, I. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. In *Language Resources and Evaluation Conference (LREC)*, pages 2126–2129, Genoa, Italy.

Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., and Rodríguez, W. R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, 51(10):948–967.

Scherer, K. R. (2000). Psychological models of emotion. In Borod, J., editor, *The neuropsychology of emotion*, pages 137–166. Oxford University Press, Oxford.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256.

Scherer, K. R., Banse, R., Wallbott, H. G., and Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15(2):123–148.

Scherer, K. R. and Ceschi, G. (2000). Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. *Personality and Social Psychology Bulletin*, 26(3):327–339.

Schiel, F., Steininger, S., and Türk, U. (2002). The smartkom multimodal corpus at BAS. In *Language Resources and Evaluation Conference (LREC)*, volume 1, pages 200–206, Las Palmas, Spain.

Schittkowski, K. (2005). Optimal parameter selection in support vector machines. *Journal of Industrial and Management Optimization*, 1(4):465–76.

Schröder, M. (2003). *Speech and emotion research*. PhD thesis, Universität des Saarlandes.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2008). Does affect affect automatic recognition of children's speech? In *Workshop on Child, Computer and Interaction*, pages 4 pages, no pagination, Chania, Greece.

Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Interspeech*, pages 805–808, Lisbon (Portugal).

Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Interspeech*, Brighton, UK.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels*. MIT Press, Massachusetts, USA.

Seppänen, T., Väyrynen, E., and Toivanen, J. (2003). Prosody based classification of emotions in spoken Finnish. In *Eurospeech*, pages 717–720, Ginebra, Suiza.

Shami, M. and Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3):201–212.

Slaney, M. and McRoberts, G. (2003). Babyears: A recognition system for affective vocalizations. *Speech Communication*, 39:367–384.

Steidl, S., Levit, M., Batliner, A., Nöth, E., and Niemann, H. (2005). Of all things the measure is man. automatic classification of emotions and inter-labeller consistency. In *ICASSP*, pages 317–320, Philadelphia, USA.

Sun, X. (2000). A pitch determination algorithm based on subharmonic to harmonic ratio. In *ICSLP*, pages 676–679, Beijing, China.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*, pages 495–518. Elsevier Science, Amsterdam.

Tato, R., Santos, R., Kompe, R., and Pardo, J. (2002). Emotional space improves emotion recognition. In *ICSLP*, pages 2029–2032.

Truong, K. P. and van Leeuwen, D. A. (2007). An 'open-set' detection evaluation methodology for automatic emotion recognition in speech. In *International workshop on Paralinguistic Speech - between models and data (ParaLing'07)*, Saarbrücken, Germany.

van Son, R. and Pols, L. (1999). An acoustic description of consonant reduction. *Speech Communication*, 28(2):125–140.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Ververidis, D. and Kotropoulos, C. (2005). Emotional speech classification using gaussian mixture models. In *IEEE Inter. Symposium on Circuits and Systems (ISCAS)*, pages 2871–2874, Japan.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007). Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. *Lecture Notes on Computer Science*, 4738:139–147.

Vogt, T. and André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME 2005).*, pages 474– 477.

Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Vogt, T. and André, E. (2009). Exploring the benefits of discretization of acoustic features for speech emotion recognition. In *Interspeech*, pages 328–331, Brighton, UK.

Wang, W., Xu, Z., Lu, W., and Zhang, X. (2003). Determination of the spread parameter in the gaussian kernel for classification and regression. *Neurocomputing*, 55:643–63.

Wendt, B. and Scheich, H. (2002). The "magdeburger prosodie-korpus". In *Speech Prosody*, pages 699–702, Aix-en-Provence, France.

Williams, C. E. and Stevens, K. N. (1981). Vocal correlates of emotional speech. In Darby, J. K., editor, *Speech evaluation in Psychiatry*, pages 189–220. Grune and Stratton, New York, USA.

Wu, K.-P. and Wang, S.-D. (2008). Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. *Pattern Recognition*.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005.

Wu, W., Zheng, T. F., Xu, M.-X., and Bao, H.-J. (2006). Study on speaker verification on emotional speech. In *International Conference on Spoken Language Processing (ICSLP)*, pages 2102–2105, Pittsburg, USA.

Xuejing, S. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *International Conference on Acoustics, Speech, and Signal Processing (ICCASP)*, volume 1, pages 333–336, Orlando, USA.

Yacoub, S., Simske, S., Lin, X., and Burns, J. (2003). Recognition of emotions in interactive voice response systems. In *Eurospeech*, pages 729–732, Ginebra, Suiza.

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., Narayanan, S., and Busso, C. (2004). An acoustic study of emotions expressed in speech. In *International Conference on Spoken Language Processing (ICSLP)*, pages 2193–2196, Jeju, Korea.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2000). *The HTK Book*. Cambridge University, Cambridge, Inglaterra.