Universidad Autónoma de Madrid

Escuela Politécnica Superior

Proyecto fin de carrera

DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO FORENSE DE LOCUTOR UTILIZANDO PARÁMETROS FONÉTICO-ACÚSTICOS Y EXTRACCIÓN AUTOMÁTICA DE FORMANTES, Y COMPARACIÓN CON EXTRACCIÓN MANUAL POR PARTE DE EXPERTOS.

Ingeniería Superior en Telecomunicación

Alberto de Castro Rodríguez

Octubre 2009

DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO FORENSE DE LOCUTOR UTILIZANDO PARÁMETROS FONÉTICO-ACÚSTICOS Y EXTRACCIÓN AUTOMÁTICA DE FORMANTES, Y COMPARACIÓN CON EXTRACCIÓN MANUAL POR PARTE DE EXPERTOS.

AUTOR: Alberto de Castro Rodríguez
TUTOR: Daniel Ramos Castro

ATVS - Biometric Recognition Group
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Octubre 2009

Agradecimientos

En primer lugar quiero agradecer a mi tutor, Daniel Ramos, el haberme brindado esta oportunidad, y especialmente por su enorme esfuerzo y dedicación y por la alegría que infunde en los que le rodeamos, que sin duda ha hecho que la elaboración de este PFC sea más agradable. y a todo el grupo ATVS por la alegría y el buen rollo que ameniza los días de los que forman parte de él. También quiero dar las gracias a Geoff Morrison, Daniel Rudoy, Yuko Kinoshita y Cuiling Zhang por sus respectivas aportaciones al desarrollo de este proyecto.

Me gustaría agradecer a mis padres, María y Carlos, y a mi hermano Carlos, por su cariño y su apoyo infatigable. Gracias a ellos me he convertido en la persona que soy hoy.

Tampoco me puedo olvidar de todos mis amigos, conocidos antes y durante la universidad que siempre le han dado sentido a todo y con los que he compartido infinidad de momentos buenos y no tan buenos.

Sin todos vosotros esto no habría sido posible, ni yo sería hoy quien soy.

Muchas gracias a todos.

Resumen

En este PFC se presentan dos sistemas de reconocimiento forense de locutor basados en características fonético-acústicas, que efectúan seguimiento automático de las trayectorias de los formantes. Cada uno se prueba sobre dos bases de datos de naturalezas diferentes.

El primero de ellos replica otro sistema en el que la extracción de formantes es realizada por un experto humano, de cara a evaluar el rendimiento ofrecido por la extracción automática. En él se extraen las trayectorias de los formantes durante pronunciaciones de diptongos o vocales, y se ajustan a curvas polinómicas o se les realizan transformaciones DCT con el objetivo de definir dichas trayectorias con un número limitado de parámetros. Estos parámetros se emplean como información discriminativa del locutor, y se emplean para obtener un valor numérico en forma de relación de verosimilitud o LR que represente el peso de la evidencia, a través de un modelado kernel de densidad multivariado (MVLR). Para cada comparación de identidades, los resultados individuales por diptongo son fusionados a través de regresión logística o suma en escala logarítmica.

El segundo sistema presenta un enfoque original, contribución de este PFC, de reconocimiento forense de locutor partiendo del sistema anterior, basado en la utilización de los diferentes parámetros (que definen las trayectorias de los formantes) como características individuales, en lugar de usarlos todos ellos de manera conjunta, permitiendo al sistema seleccionar e integrar tan sólo aquellos parámetrosque ofrezcan un mejor rendimiento siguiendo una estrategia de selección de características.

Ambos sistemas se probarán sobre dos bases de datos diferentes: *Kinoshita & Osanai 2006* de hablantes de inglés australiano y obtenida en condiciones controladas (habla microfónica controlada etc.), y *Zhang 2007* de hablantes de chino mandarín y obtenidas en condiciones más duras (codificación GSM, habla espontánea etc.).

Por último se realiza un análisis general del sistema y su funcionamiento para la obtención de conclusiones finales y el planteamiento del posible trabajo futuro relacionado con este proyecto.

Una parte de este proyecto ha sido aceptado y publicado en la 10^{th} Annual Conference of the International Speech Communication Association (ISCA): Interspeech 2009 Brighton (http://www.interspeech2009.org), congreso internacional de máximo impacto en el área de tratamiento de voz, en el artículo [1] reproducido en el Apéndice A:
A. de Castro, D. Ramos, and J. Gonzalez-Rodriguez, *"Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking"* in *Proc. of Interspeech*, 2009.

Palabras Clave

Reconocimiento forense de locutor, características fonético-acústicas, seguimiento automático de formantes, ajuste paramétrico, selección de características, relación de verosimilitud.

Abstract

This volume presents two different forensic speaker recognition systems based on acoustic-phonetic features, that use automatic formant-trajectory tracking. Both of them are tested on two different-nature databases.

The first one replicates another system in which formants are manually tracked by human experts, so that the performance of the automatic formant tracking can be evaluated directly by comparing results. The formant trajectories are tracked over diphthong or vowel pronunciations, and are fitted to polynomials or DCT transformed in order to resume the whole trajectory into a reduced amount of parameters. Those parameters are taken as locutor-discrimative information, and are used to generate a numeric value called likelihood ratio with a Multi-Variate Kernel Density model (MVLR). For each identity comparison the individual per-diphthong results are fused applying logistic regression or direct logarithmic-scale adding.

The second system presents a new approach for forensic speaker recognition, as an original contribution, based on the use of the different parameters that define the formant trajectories as individual features that contain locutor information, instead of using all of them together as a single package. In this way, the system is allowed to select and integrate only the individual parameters that improve the system performance, following a feature selection strategy.

Both systems will be tested on two different databases in order to evaluate their performance: *Kinoshita & Osanai 2006* which contains records for Australian-English speakers under controlled conditions (microphonic recording, control sentences etc.), and *Zhang 2007* which contains recordings for Mandarin Chinese speakers in tough conditions (GSM codification, spontaneous speaking etc.).

In last term, a general analysis is made over the system results and working, in order to obtain final conclusions and define possible future work related wit this project.

Part of this project has been accepted for publishing in the 10^{th} Annual Conference of the International Speech Communication Association (ISCA): Interspeech 2009 Brighton (http://www.interspeech2009.org), an international congress with maximum impact in voice treatment. The article [1] is reproduced in the Appendix A:
A. de Castro, D. Ramos, and J. Gonzalez-Rodriguez, *"Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking"* in *Proc. of Interspeech*, 2009.

Key words

Forensic speaker recognition, acoustic-phonetic features, automatic formant tracking, parametric fitting, feature selection, likelihood ratio.

Glosario de acrónimos

- **APE**: Applied Probability of Error
- **DCT**: Discrete Cosine Transform
- **DET**: Detection Error Trade-Off
- **EER**: Equal Error Rate
- **FA**: False Accept
- **FR**: False Reject
- **FSR**: Forensic Speaker Recognition
- **F**X: Formante(s) X
- **GSM**: Global System for Mobile
- **JPEG**: Joint Photographic Experts Group
- **LPC**: Linear Predictive Coding
- **LR**: Likelihood Ratio
- **MVKD**: Multi-Variate Kernel Density
- **MVLR**: Multi-Variate Likelihood Ratio
- **PAV**: Pool Adjacent Violators
- **PCM**: Pulse-Code Modulation
- **VAD**: Voice Activity Detection

Índice de Contenidos

Índice de Figuras

Índice de Tablas

1
Introducción

1.1. Preámbulo

A lo largo del desarrollo de un juicio, para ayudar a determinar la implicación o no en los hechos de un individuo presuntamente involucrado, se puede hacer uso del reconocimiento forense de locutor, manifestado en el apoyo a una de las dos hipótesis que simbolizan los intereses de fiscalía y la defensa respectivamente, especialmente en sistemas adversariales (modelo anglosajón). Para poder recurrir al reconocimiento forense de locutor, debe ser presentada como prueba por alguna de las partes una grabación de legitimidad legal contrastada que influya en la determinación de la culpabilidad o inocencia del individuo en cuestión (por ejemplo una grabación policial generada por el agresor). Su comparación con una serie de muestras de control tomadas de un sospechoso es lo que se conoce como *evidencia*. Interesa por tanto en estos casos valorar el peso de la evidencia para determinar su relevancia en el proceso. Para ello, se obtienen también de manera controlada muestras de referencia del sospechoso, haciéndole pronunciar unas frases o palabras determinadas y recogiéndolas en un ambiente conocido, con el objetivo de recopilar la información necesaria para poder llegar a conclusiones concretas sobre el asunto que nos ocupa.

En este proyecto se presenta un sistema automático de reconocimiento forense de locutor, que persigue el objetivo de evaluar el peso de la evidencia a través de la comparación entre una locución incriminatoria, de identidad desconocida, y una locución de origen conocido. El reconocimiento forense de locutor [2] engloba áreas multidisciplinares como la lingüística, la fonética, la acústica, el procesado de señal o la estadística, yendo mucho más allá del *reconocimiento nativo* [3] de locutor, entendido como la habilidad innata de los seres humanos para identificar al individuo que genera una locución que llega a sus oídos.

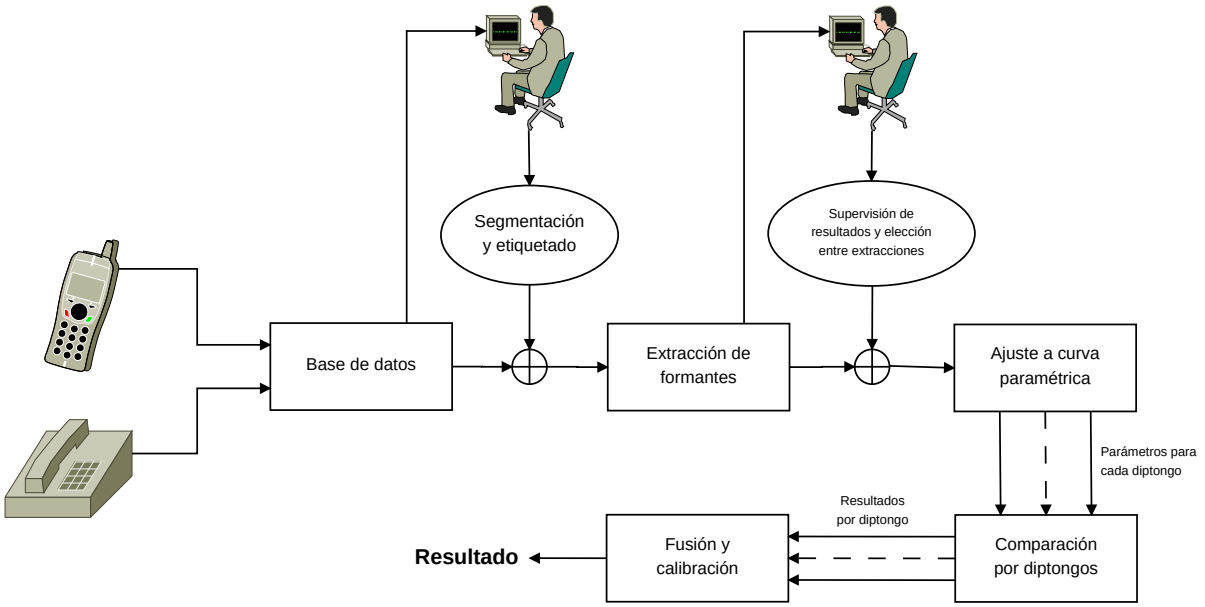


Figura 1.1: Esquema de un sistema forense de reconocimiento de locutor basado en ajuste paramétrico con extracción de formantes semi-automática

A día de hoy existen, y se trabaja, en múltiples formas [4] de afrontar el problema que plantea el identificar una locución dubitada, entendida como aquella cuya identidad nos es desconocida. Pero la mayoría de ellas están basadas bien en procedimientos automáticos de reconocimiento de locutor ([5, 6, 7] etc.) o bien en aproximaciones basadas

en análisis humano de características fonético-acústicas, también llamadas tradicionales [8]. La mayoría de métodos optan por una de estas dos aproximaciones o una combinación de ambas para desarrollar sistemas de reconocimiento de locutor.

Por ejemplo en la Figura 1.1 se resumen las fases del sistema forense de reconocimiento de locutor basado en características fonético-acústicas descrito en [9]. Se trata de un sistema completo, que emplea una base de datos que nos es disponible, y para el que se dispone de los resultados generados sobre dicha base de datos. A continuación se desarrollan brevemente las diferentes fases:

1. **Base de datos**, generadas a partir de la captación de un conjunto de locuciones en condiciones generalmente similares, de tal forma que sirvan para efectuar pruebas de rendimiento de sistemas, o puedan ser usadas como información relevante de una población tipo.

2. **Segmentación y etiquetado** realizados por un experto humano, consisten en la delimitación temporal de la unidad lingüística de interés en cada locución, y la asignación de una denominación que defina dicha unidad, de cara a compararla únicamente con unidades similares o compatibles en otras locuciones.

3. **Extracción de formantes** de forma semi-automática, para cada región analizada, se extraen de manera automática las trayectorias de los formantes. Se generan varias posibles extracciones siguiendo diferentes criterios, y el experto humano decide en la siguiente etapa.

4. **Supervisión de resultados y elección entre extracciones** realizadas por experto humano, que comprueba la validez de las extracciones semi-automáticas, y elige entre ellas la mejor en base a comprobaciones visuales o escucha de sonidos sintetizados a partir de la información extraída.

5. **Ajuste a curva paramétrica** de las trayectorias de formantes extraídas de tal forma que estas queden lo mejor definidas posible por un número reducido de parámetros. Se emplea ajuste polinómico o transformación DCT (Discrete Cosine Transform).

6. **Comparación por diptongos:** empleando estos parámetros para evaluar el parecido entre diferentes muestras de diferentes usuarios en base al conocimiento previo proporcionado por las grabaciones del resto de usuarios en la base de datos. Para cada comparación se genera un resultado diferente en base a cada diptongo comparable.

7. **Fusión y calibración** para cada comparación. Los resultados generados por los diferentes diptongos son fusionados en un único resultado y este es calibrado con el objetivo de alcanzar un único resultado final que resuma lo mejor posible el parecido entre las identidades comparadas.

En cambio la Figura 1.2 muestra el esquema de otro sistema en el que la extracción semi-automática de formantes ha sido sustituida por un método de extracción plenamente automático, lo que hace prescindible el esfuerzo del experto forense en esta tarea, que puede así ser efectuada íntegramente sin precisar atención humana. Esto allana mucho el camino hacia el uso de grandes bases de datos, lo que siempre es deseable tanto para obtener LRs más fiables, como para una validación estadísticamente robusta.

Este sistema es presentado en la Sección 3.4 de este proyecto para evaluar la bondad de la extracción automática de formantes, para la que se hace uso de la herramienta

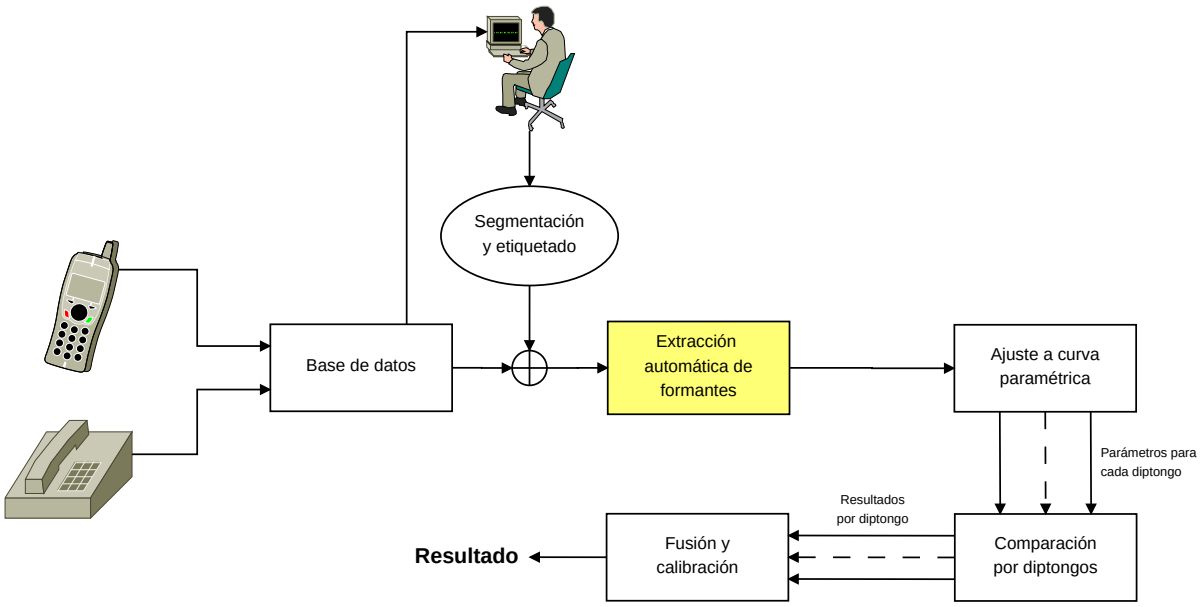


Figura 1.2: Esquema de un sistema forense de reconocimiento de locutor basado en ajuste paramétrico con extracción de formantes plenamente automática

para MATLAB^{TM} descrita en [10]. El resto del sistema es en esencia una réplica del anterior, intentando replicar el resto del experimento en condiciones similares para obtener una evaluación fiable del detrimento en el funcionamiento del sistema que acarrea la automatización. Para evaluar esta reducción del rendimiento, se comparan los resultados obtenidos por este sistema con los resultados del sistema con extracción semi-automática de formantes, reflejados en [9].

Adicionalmente se presenta un sistema basado en un enfoque nuevo, que mejora el rendimiento aprovechando el diferente poder discriminativo de los parámetros fonético-acústicos individuales de cada diptongo y formante en la base de datos. La mejora radica en el uso individualizado de las características, distinguiendo y evaluando cara parámetro individual del conjunto que define la trayectoria de un formante, en lugar de emplear todos ellos de manera conjunta.

La Figura 1.3 muestra el esquema del sistema basado en selección de características que se presenta en este proyecto en la Sección 3.5. Dicho sistema se sirve de la extracción de formantes y el ajuste paramétrico descrito en el sistema anterior.

Sin embargo, no evalúa todos los parámetros que definen las trayectorias de los formantes de manera conjunta, sino que considera cada uno de ellos como una característica del locutor en sí misma, y recurre a una discriminación paramétrica de acuerdo con el rendimiento ofrecido por cada uno de los elementos disponibles.

De tal modo, selecciona secuencialmente aquellas características que ofrecen un nivel de rendimiento suficientemente alto como para mejorar el rendimiento global del sistema, que solo puede mejorar con la adición de cada nueva característica.

A lo largo de este Capítulo se presenta la motivación para el desarrollo de este proyecto, los objetivos perseguidos por el mismo en base a dichas motivaciones, y una descripción general de los sistemas desarrollados que se presentan en esta memoria. Adicionalmente se describe la estructuración global de la misma.

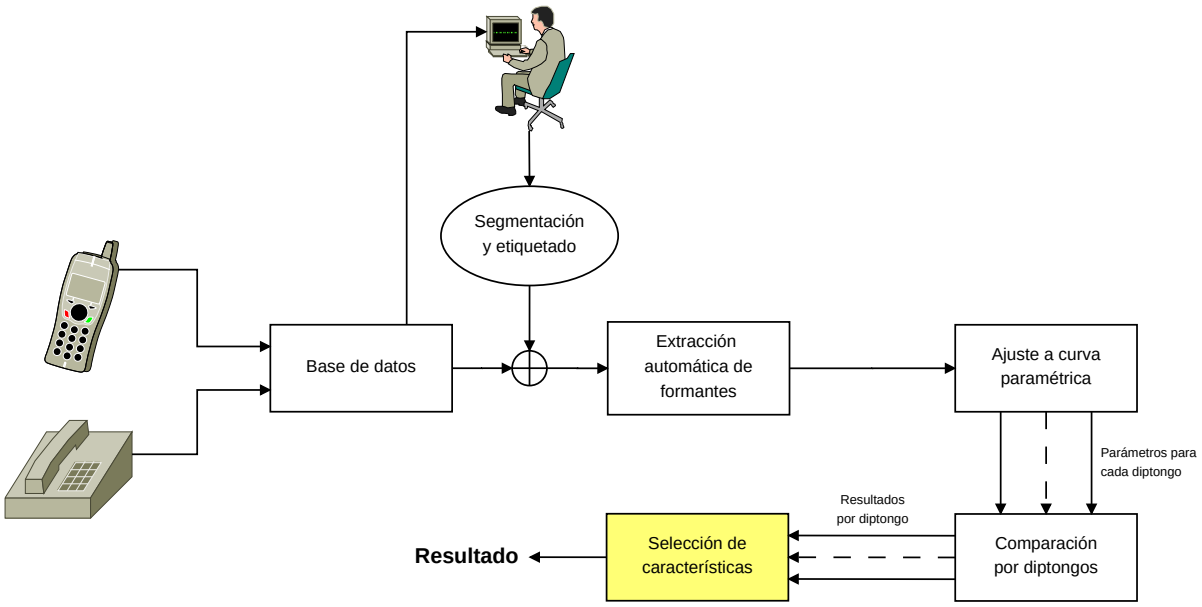


Figura 1.3: Esquema de un sistema forense de reconocimiento de locutor basado en ajuste paramétrico con extracción de formantes plenamente automática

1.2. Motivación

El reconocimiento forense de locutor puede presentar una serie de limitaciones. Un sistema de reconocimiento forense de locutor generalmente precisa de un conjunto de locuciones suficientemente representativas de la población tipo a la que pertenecen el agresor y el sospechoso, que necesariamente debe ser la misma para que puedan tratarse del mismo individuo. La escasez de datos de voz en poblaciones afecta de manera directa al rendimiento de los sistemas de reconocimiento forense de locutor.

Además es probable que cada una de estas muestras requiera un cierto tratamiento previo. Muchos sistemas contemplan un tratamiento manual de las muestras de voz (en forma de normalización, segmentación, etiquetado, rechazo etc. de las muestras). Esto se convierte en un problema cuando se pretende trabajar con conjuntos de muestras muy amplios, en los que la manipulación de la totalidad de muestras deja de ser viable (por ejemplo no es razonable pretender etiquetar y segmentar manualmente millones de muestras).

Este hecho hace que sea deseable automatizar al máximo posible los sistemas de reconocimiento forense de locutor, de tal manera que generen resultados en márgenes de tiempo razonables, y sean aplicables sobre bases de datos de grandes dimensiones lo que conlleva grandes ventajas como el aumento de generalidad al poder aplicarse sobre bases de datos nuevas, y una mayor robustez estadística de los resultados al tener en cuenta mayor cantidad de muestras de referencia de la población.

1.3. Objetivos

A lo largo de este proyecto se persiguen los siguientes objetivos:

1. La automatización de la extracción de la máxima información posible sobre la identidad del locutor de una grabación bajo estudio sirviéndose de una base de datos de locutores conocidos generada por expertos forenses.

2. Reducir o eliminar el esfuerzo humano experto implícito en el manejo de bases de datos de grandes dimensiones, principalmente aquel requerido en la labor de extracción de formantes y en el tratamiento previo de señal, buscando la posibilidad de manejar una mayor cantidad de datos de entrada, lo que se traduce en una mejora sustancial de resultados.

3. Evaluar el coste implícito por el aumento del grado de automatización del sistema, en forma de una posible reducción en el rendimiento de los resultados, comparando estos con otros generados por otros sistemas con menor grado de automatización que requieren el trabajo o la supervisión de expertos fonetistas para su correcto funcionamiento, siempre que esta comparación sea posible.

4. Analizar los parámetros fonético-acústicos (por ejemplo trayectorias de formantes en diferentes diptongos) y su influencia en el rendimiento con el fin de obtener información relevante acerca del poder de cada característica para discriminar locutores, de manera que sea posible valorar el rendimiento de cada una de ellas.

5. Apoyándose en esta información, mejorar en lo posible el rendimiento de los sistemas forenses de reconocimiento de locutor basándose en parámetros fonético-acústicos, evaluando la información aportada por cada una de sus características y aprovechando aquellas más discriminativas y/o aquellas que mejor influyan en el funcionamiento global del sistema.

1.4. Contribuciones originales

Una de las principales contribuciones de este proyecto es el incremento del grado de automatización frente a otros sistemas de reconocimiento de locutor. Esta mayor automatización se concreta principalmente en la supresión del esfuerzo humano requerido en la labor de la extracción de formantes y en el tratamiento previo de señal, que tradicionalmente ha requerido de un experto fonetista para su supervisión o incluso realización completa.

Otra contribución es la presentación de una estrategia basada en selección de características, posterior al ajuste paramétrico, que discrimina los parámetros de acuerdo a su rendimiento individual. Se seleccionan e integran aquellas características que suponen una mejora en el funcionamiento del sistema y por el contrario se descartan aquellas que disminuyen la calidad de los resultados.

1.5. Organización de la memoria

Este proyecto está organizado como se describe a continuación. En el Capítulo 1 se hace una introducción global al proyecto, se presentan la motivación y los objetivos del proyecto, con un breve resumen de las contribuciones originales incluidas en el proyecto y una descripción de la organización de la memoria. El Capítulo 2 contiene una revisión del estado del arte y los trabajos relacionados que pueden resultar interesantes para una mejor ubicación y comprensión de este proyecto. A lo largo del Capítulo 3 se describe en detalle cada fase de los sistemas desarrollados, en primer lugar se describen las fases comunes, y a continuación se especifican las fases propias de cada uno de ellos. El Capítulo 4 contiene la parte experimental del proyecto, dónde se demuestra la adecuación de la extracción automática de formantes, se efectúa una valoración crítica de su rendimiento y se compara con extracción por parte de expertos humanos. Se efectúa también una extensa

comparación con otros sistemas, todo ello sobre dos bases de datos muy diferentes. En el Capítulo 5 se exponen las conclusiones que se extraen del desarrollo del proyecto y especialmente de los resultados reflejados en el Capítulo 4, además del posible trabajo futuro a realizar a partir del mismo.

2

Estado del arte y trabajos relacionados

2.1. Introducción al Capítulo

Históricamente el reconocimiento forense de locutor basado en parámetros fonético-acústicos se ha afrontado de diferentes formas. Aun a día de hoy no existe un método universal con el que obtener información del locutor de una evidencia en forma de habla grabada, sino que se sigue investigando por diferentes vías y comparando los resultados obtenidos por diferentes métodos, que año a año son mejores.

A lo largo de este Capítulo se describirán los diferentes tipos de información contenidos en grabaciones de voz, con la intención de definir la información fonético-acústica. También se revisarán algunos métodos de extracción de características fonético-acústicas, especialmente de trayectorias formánticas en las que se basa este proyecto, y se plantean los retos del reconocimiento forense de locutor.

También se detallará la metodología de expresión de resultados en forma de Relaciones de Verosimilitud (Likelihood Ratios - LR) que será la base para realizar la comparación entre las características de habla dubitada e indubitada, y se explicarán diversas medidas de rendimiento, y representaciones gráficas que reflejen el funcionamiento de los sistemas. Adicionalmente se entrará en detalle sobre el método semi-automático de extracción de formantes planteado y utilizado en [9], trabajo sobre el cual se desarrolla parte del siguiente proyecto, y que requiere la participación de expertos humanos.

2.2. Características del habla discriminativas por locutor

Si nos disponemos a extraer y analizar información de grabaciones de voz, en primer lugar se deben conocer las características del habla inherentes al locutor y que por tanto pueden ser sujeto de estudio, y la relevancia de las mismas. Entonces se procederá al estudio del parecido de dichas características entre grabaciones de locutor desconocido (y cuya identidad queremos averiguar) y grabaciones de control tomadas sobre un individuo sospechoso de ser dicho locutor. Ambas se comparan en el contexto de una población tipo.

Antes de emitir un resultado final, se debe tener en cuenta no sólo el parecido en una o varias de las características analizadas, sino la tipicalidad de las mismas en la población. Un conjunto de características similares entre un locutor dubitado y uno indubitado puede llevar fácilmente a apoyar la hipótesis de que ambos sean el mismo individuo, sin embargo antes de emitir una valoración al respecto se debe tener en cuenta el hecho de que esta distribución de características pueda ser bastante común entre la población tipo, por lo que la teoría de la fuente común pierde fuerza, o por el contrario puede tratarse de un conjunto de características atípico y cuya repetición en otro individuo sea altamente improbable.

Las características propias del locutor de una grabación lo distinguen en mayor o menor medida (según la tipicidad) de otros locutores, aunque cabe destacar que la mayoría de estas propiedades no son constantes. Es más, aquellas que son medibles ni siquiera toman valores discretos, sino que generalmente vienen descritas por funciones de densidad de probabilidad [11], que definen cómo se distribuyen las probabilidades de las características en locuciones de un individuo particular. Esto se traduce en que el mismo locutor generará diferentes valores en diferentes momentos o incluso en realizaciones consecutivas (variabilidad intra-locutor). Esta diferencia no será tampoco la misma en unas personas que en otras. En condiciones óptimas será sensiblemente menor que la variación frente a muestras de otros individuos (variabilidad inter-locutor).

Las características propias de un locutor pueden clasificarse según su naturaleza como la intersección de dos clasificaciones binarias según [8]. La primera de ellas permite diferenciar entre características auditivas y características acústicas:

- **Características auditivas:** son características auditivas aquellas que pueden ser percibidas directamente en la escucha por un oyente entrenado en estructura lingüística, limitado no solo a fonética. Dicho entrenamiento consistiría, por un lado en transcribir e interpretar cualquier locución conversacional (fonética), pero también en analizar la estructura lingüística, y como ésta cambia en un mismo usuario y entre usuarios distintos. Las diferencias se perciben no como el resultado de sub-análisis de diferentes características, sino de manera global en la escucha, y se expresaría como la sensación de similitud transmitida, aunque a ésta podría asignársele un valor numérico en forma de relación de verosimilitud [12].

- **Características acústicas:** son aquellas que requieren alguna manipulación de señal para ser evaluadas, no son percibidas directamente en la escucha. Las características acústicas se subdividen a su vez entre tradicionales y automáticas. Las características *tradicionales* son aquellas relacionadas directamente con la producción de la voz, tales como frecuencias formantes, frecuencia fundamental etc. Las características *automáticas* se extraen de forma indirecta, como en el caso de los coeficientes cepstrales o delta-cepstrales, descritos en [13, 14]. En ocasiones alcanzan un rendimiento superior al de las características tradicionales, sin embargo presentan una contrapartida bastante importante: los valores numéricos que toman las características automáticas no son directamente interpretables, lo cual dificulta su explicación en un juicio. A pesar de ello, no existe una razón para que características automáticas y tradicionales no puedan utilizarse conjuntamente en forma de datos multivariables.

En general se recurre al uso conjunto de características auditivas y acústicas. Las características acústicas son más rigurosas a la hora de generar un resultado, sin embargo puede darse el caso de que dos locuciones que generan unos valores acústicos cercanos nos lleven a pensar que vienen de la misma fuente, sin embargo un análisis auditivo (una escucha) lo descarte directamente, al encontrar alguna característica muy diferente entre ambas muestras que no haya repercutido en los resultados de los análisis acústicos realizados. El análisis auditivo también se antoja imprescindible, por ejemplo, en la segmentación de conversaciones formadas por varios individuos y la identificación de cada uno de ellos (aunque resulten desconocidos) en los fragmentos en que participan.

Una segunda clasificación binaria descrita en [8], independiente de la anterior, diferencia entre características lingüísticas y características no-lingüísticas según el criterio descrito a continuación:

- **Características lingüísticas:** las características lingüísticas son aquellas relacionadas con el orden y la estructuración de las diferentes unidades del lenguaje. Principalmente se pueden categorizar en:

    1. **Fonológicas:** aquellas relacionadas con realizaciones alofónicas de fonemas. Los alófonos son sonidos del habla, realizaciones diferenciadas de un mismo fonema (por ejemplo en la palabra *dado* generalmente pronunciamos /daðo/). Así, dos individuos diferentes pueden utilizar alófonos diferentes para pronunciar el mismo texto (llegando a extremos como el ceceo, el seseo etc.).

2. **Morfológicas:** aquellas relacionadas con la estructuración interna de las palabras, basada en los morfemas, entendidos estos como la unidad mínima del lenguaje con significado propio. Los alomorfos son las diferentes realizaciones fónicas de un determinado morfema. Por ejemplo, en español el diminutivo masculino se forma como -ito pero también como -illo, -ín, -ino etc. Cada locutor puede presentar características personales, contemplándose el acento, la pronunciación o la idiosincrasia entre otras, así como otros dependientes del nivel cultural o dialectal.
3. **Sintácticas:** aquellas relacionadas con la coordinación y la forma en que se juntan y organizan las palabras en estructuras superiores como sintagmas u oraciones para expresar conceptos complejos. Individuos diferentes pueden construir oraciones diferentes para expresar la misma idea.

- **Características no-lingüísticas:** son todas aquellas que no se pueden encuadrar como lingüísticas. Son generalmente fruto de problemas articulatorios o fonatorios, como por ejemplo una voz apagada, un timbre nasal, o una dicción lenta. Pueden tener origen patológico, desde afonías o constipados hasta defectos congénitos.

Generalmente los desarrollos de sistemas automáticos, no aprovechan toda la información relevante del locutor al no estudiar todas las características propias conocidas, debido a varios motivos como por ejemplo el acarreo de un tiempo excesivo implicado o la inviabilidad de la automatización, etc. por lo que diferentes métodos pueden generar resultados distintos, aunque estos pueden ser complementarios. El enfoque a la hora de integrar diferentes tipos de información puede partir de planteamientos diferentes, desde el desarrollo de un sistema multinivel que contemple varios tipos de información hasta el uso de multisistemas que se sirvan de una diversidad de resultados generados de forma aislada. En [15] se desarrolla un sistema multimodal que emplea información relativa al tracto vocal, combinada con información basada en el uso del lenguaje a alto nivel.

2.3. Uso de frecuencias formantes en reconocimiento forense de locutor

Los formantes, son máximos en la transformación frecuencial de la señal debidamente enventanada durante la pronunciación de determinados fonemas. Debido a las singularidades de cada tracto vocal, y al aprendizaje individualizado del lenguaje, cada individuo presenta unos valores característicos de frecuencias formánticas y trayectorias temporales de estas, así como propia de dichos formantes, durante la pronunciación de sonidos concretos.

El uso de frecuencias formantes implica una serie de ventajas frente a otro tipo de características fonético-acústicas, que hacen que su uso en sistemas de reconocimiento forense de locutor sea apropiado, tales como su definición absoluta en forma de valor numérico, y que actualmente ya haya sido empleado en procesos judiciales reales, gracias además a ser fácilmente explicable a personas no versadas en la materia tales como jueces, jurados etc.

La trayectoria seguida por dichos formantes a lo largo de realizaciones concretas de un fonema o grupo de fonemas, puede ser empleada para el reconocimiento forense de locutor. Comparando los comportamientos de los mismos, extraídos de diferentes grabaciones, se puede obtener cierta información relativa a la identidad del locutor de la muestra bajo análisis cuya identidad es desconocida.

Sin embargo la extracción de formantes es una tarea no trivial, y todavía no está resuelta en la literatura [16]. Los mejores resultados se obtienen cuando expertos se dedican a extraerlos manualmente de las regiones de interés, no obstante para un gran número de grabaciones, cada una de ellas con un alto número de ventanas que precisan la extracción individual de formantes para cada una de ellas, es una tarea que consume una gran cantidad de tiempo.

2.3.1. Extracción semi-automática de formantes

El procedimiento descrito en [17] (basado en la aplicación práctica del algoritmo que se recoge en [18]) que genera valores *candidatos* para los tres primeros formantes a partir de coeficientes de autocorrelación LPC y a un valor máximo de corte para F3. Si se generan exactamente 3 candidatos se asignan directamente a F1 F2 y F3, en caso contrario se elige entre los posibles candidatos la opción que genere una menor discontinuidad con respecto a las ventanas adyacentes.

Las tasas de resultados para varones adultos, con un margen de error de 300Hz (que no deberían suponer omisiones de ningún formante) para F1 F2 y F3 son de 100 %, 99 % y 94 % respectivamente, presentando unos errores cuadráticos medios de 63Hz, 96Hz y 172Hz respectivamente. Se observa que la extracción de F3 ofrece un menor rendimiento.

En [18] se refleja que en el 85-90 % de las extracciones se obtienen 3 candidatos razonables por debajo de los 3kHz para los valores de los 3 primeros formantes a partir de la autocorrelación LPC (*Linear Predictive Coefficients*), empleando LPC de orden 15, y frecuencia de muestreo de 10kHz. También se obtuvieron buenos resultados para mujeres y niños adaptando la frecuencia de muestreo y el umbral máximo para F3 a valores apropiados para cada locutor. El rendimiento con niñas es sensiblemente inferior.

En [19] se presenta un procedimiento semi-automático para la extracción de formantes que hace uso del procedimiento anterior de cara a acelerar la extracción de formantes. Parte de una segmentación manual de las regiones de interés sobre los ficheros de audio. Se efectuaron medidas según el procedimiento anterior cada 2ms con ventanas temporales de 100ms, utilizando una ventana cos^4. Se efectuaron ocho mediciones de los formantes con diferentes umbrales para F3 entre 2500Hz y 4000Hz, obteniendo ocho conjuntos de trayectorias candidatas para los formantes.

Para cada conjunto de valores se podía escuchar una sintetización artificial de la vocal junto con la grabación original, además de una representación gráfica de las trayectorias superpuestas al espectrograma original de la señal. Con ayuda de esta inspección visual y auditiva, el experto elige de forma manual entre los diferentes candidatos generados.

2.4. Retos en el rendimiento de las técnicas de reconocimiento forense de locutor basado en características fonético-acústicas

En esta sección se enumeran una serie de retos o problemas que plantea el reconocimiento forense de locutor basado en características fonético-acústicas. A día de hoy la forma de afrontarlos en mayor o menor medida evoluciona con la investigación y desarrollo de la tecnología.

2.4.1. Extracción de características

La extracción de características es uno de los campos más abiertos, y más cambiantes, si englobamos en este campo la selección de las características que se extraen, por ejemplo en [20] se evalúan dos características diferentes. Es un campo de vital importancia. Un extractor ideal de un número de características altamente discriminativas, generaría unos resultados excelentes con un sistema discriminador básico. Sin embargo para la mayoría de extractores y características reales se requiere la sofisticación del sistema discriminador para ofrecer resultados aceptables, debido a que la información aportada por las características individuales no es plenamente discriminativa.

2.4.2. Ruido y distorsión

Hablamos de ruido y distorsión, no entendido únicamente como impurezas en grabaciones, sino como la aparición de cualquier fenómeno que afecte las medidas que se efectúan sobre las muestras de interés y que no provenga de características subyacentes de las mismas. La causa del ruido puede ser cualquier sonido no deseado en el momento de la grabación, un defecto en el canal, la degradación con el tiempo de una unidad de almacenamiento etc. [21]. La aparición de cualquier tipo de ruido o distorsión reduce la fiabilidad de las medidas de características, y por lo tanto el rendimiento del sistema.

En casos reales, las grabaciones disponibles presentan siempre ciertos niveles de ruido, debido por ejemplo a la ocultación de un micrófono en un cajón, limitación frecuencial en el canal telefónico, que la voz haya sufrido una codificación, por ejemplo GSM, o cualquier fuente de ruido en el ambiente de la grabación.

2.4.3. Modelado de poblaciones

Es imprescindible disponer de alguna información previa para elaborar un sistema de reconocimiento, para conocer que características son analizables sobre las muestras. A la hora de emitir una valoración también es fundamental conocer la distribución de estas características sobre la población a la que pertenecen el agresor y el sospechoso, aunque es posible obtener dicha distribución a partir de un conjunto de muestras representativo en la fase de entrenamiento previo del sistema. Para ello es necesario disponer de bases de datos extensas tales como [20, 22, 23].

2.4.4. Selección de unidades fonéticas

El criterio de segmentación es una de las facetas más importantes a la hora de desarrollar un sistema de reconocimiento de locutor. Se pueden tomar diferentes decisiones respecto a la unidad básica del estudio, y analizar individualmente fonemas, diptongos, sílabas, palabras, o cualquier estructura que se desee.

Por ejemplo en [23] se describe un sistema de reconocimiento de locutor basado en el estudio de los formantes, que toma como unidad básica de estudio fonemas, mientras que en [24] se habla del poder discriminatorio de los diptongos y en [25] se detalla un experimento completo basado en un único diptongo.

2.4.5. Variabilidad

Como ya se ha comentado anteriormente, existen características extraíbles de habla grabada que son discriminativas del locutor, y pueden ser muy útiles para diferenciarlo de otros locutores. Sin embargo hay que tener en cuenta, que para el mismo individuo, este

valor no siempre permanecerá constante en distintas observaciones, aunque sí se espera que no oscile demasiado, pues en ese caso dejaría de ser discriminativa.

Existe un amplio conjunto de factores que pueden influir en estas características, desde el estado de ánimo hasta el sistema empleado para capturar el habla, y aunque se replicara en igualdad de condiciones, no se garantizaría la repetición del mismo valor, pues pueden tomar infinitos valores al tratarse de magnitudes continuas, aunque quedan definidas por una función de densidad de probabilidad.

La variabilidad afecta al rendimiento del sistema. En sistemas automáticos es un fenómeno muy tratado, por ejemplo en [26] se especifica un modelado de variabilidad inter-sesión y en [27] se propone un método que tiene en cuenta efectos de canal, en ambos casos para sistemas de verificación de locutor

2.4.6. Escasez de muestras

En general la cantidad y la calidad de los conjuntos de muestras forenses disponibles son reducidas, tal y como se relata en [28, 5]. Esta circunstancia afecta de manera directa al rendimiento del análisis forense, debido esencialmente a lo reducido del número de muestras es más difícil modelar las diferencias sustanciales entre las muestras de habla de identidad dubitada, muestras de control de sospechosos, y muestras representativas de la población.

Esto repercute directamente en el modelado de la variabilidad inter-sesión de los usuarios dificultándolo significativamente, pues la escasez de muestras y la variación externa entre ellas y con respecto a las muestras de control y de la población, impide efectuar una evaluación realista de la misma, lo que afecta de manera importante a los resultados finales de las comparaciones.

2.5. Interpretación de evidencias forenses

Cuando un experto o un sistema, analizan una evidencia relativa a un procedimiento judicial, buscan sacar conclusiones valiosas para el desarrollo del proceso. No obstante en ningún caso serán responsables de emitir una sentencia, o una conclusión final sobre la culpabilidad o inocencia de los locutores, estas tareas corresponden a otros individuos (jueces, jurados etc.) [5, 29]. Por tanto el rol del perito forense de se debería limitar a la evidencia y a expresar de la mejor forma posible la información relevante relativa a la misma en relación con la población manejada. Sin embargo, la forma de expresar esta información, no es ni mucho menos obvia.

2.5.1. Expresión de resultados del análisis forense

No existe a día de hoy una forma universal de emitir evaluaciones acerca de la comparación entre muestras forenses en un informe. Sin embargo sí existen algunos conceptos que se empiezan a adoptar de forma general, como evitar afirmaciones categóricas, sino más bien orientar la valoración del experto forense de tal forma que sirva de asesoramiento a jueces, jurados etc. para que decidan por sí mismos, evitando resultados que sean conclusiones finales en sí mismos.[30, 31]

En [32] se propone un esquema para la expresión de conclusiones de reconocimiento de locutor realizada por expertos, con la intención de unificar criterios a nivel nacional en el Reino Unido, en forma de dos clasificaciones por niveles, basadas no sólo en la diferencia de características entre muestras sino también en la tipicalidad de los valores.

En una primera clasificación, se pretende evaluar la diferencia entre muestras, de cara a decidir si ambas muestras son compatibles, es decir, consistentes con haber sido producidas por el mismo locutor, analizando las diferencias entre ambas y evaluando con que fuerza pueden ser fruto de la variabilidad inter-sesión. Por tanto se puede concluir clasificando la evidencia como:

1. Consistente
2. No consistente
3. Sin decisión

Sólo en el caso de que se haya concluido que las muestras son consistentes con haber sido producidas por el mismo locutor, se procede a una segunda clasificación, basada en la distintividad de las características comunes con respecto del grueso de la población. Dicha distintividad se evalúa sobre una escala de cinco puntos:

1. No distintivo
2. Moderadamente distintivo
3. Distintivo
4. Altamente distintivo
5. Excepcionalmente distintivo (la posibilidad de que la combinación de características se repita en otro individuo es remota).

Se considera un caso especial el reconocimiento de locutor sobre un conjunto cerrado de individuos (por ejemplo, los participantes en una conversación registrada, las personas en una ubicación de acceso controlado etc.). Si existen diferencias sustanciales entre las características contempladas, se justifica una identificación categórica.

Sin embargo esta forma de expresión de resultados del análisis de la evidencia no está aceptada globalmente, ni mucho menos, sino que han aparecido varias respuestas discordantes. Por ejemplo en [33] se critica que la comparación de dos muestras extremadamente parecidas pueda resumirse en un término que no transmita esta situación en caso que los valores de la característica analizada no sean altamente atípicos. También discrepa en el uso de dos escalas según consistencia y distintividad no relacionadas de manera directa, de tal forma que no es posible expresar si las diferencias entre las muestras se deben presumiblemente a la variabilidad inter-sesión, a fuentes diferentes para cada grabación, o ambas posibilidades son igualmente probables. También se critica que el uso de dos fases de evaluación no es representativo de la práctica forense moderna, y que las posibles clasificaciones generen necesariamente un número finito de categorías, siendo necesario por tanto el uso de umbrales y dándose por tanto un efecto cliff-edge (ver [34]) para muestras cercanas a ellos, es decir, que dos puntuaciones casi idénticas pueden resultar en dos categorías próximas con nomenclaturas de alta diferencia semántica. También denota la ausencia de una especificación acerca del manejo de datos multivariados, y que las comparaciones sobre conjunto cerrado pueden ser tratadas exactamente de la misma manera que sobre conjunto abierto desde el puto de vista del peso de la evidencia tal y como se demuestra en [2]. También discrepa en detalles como la elección de la terminología etc.

2.5.2. Inferencia bayesiana de la identidad en reconocimiento forense de locutor

La labor del experto debe limitarse a su ámbito, se reduce a la evidencia recogida y a la información técnica de la misma que le sea proporcionada, y en base a ello deben ser elaboradas las conclusiones que extraiga. Un modelo probabilístico basado en el teorema de Bayes puede resultar muy útil de cara a:

1. Asistir a los expertos a la hora de evaluar el peso de la evidencia, tanto para generar resultados como para expresar los mismos de una forma inteligible para personal ajeno a la investigación.

2. Ayudar a los juristas y a personal no experto en reconocimiento forense de locutor a interpretar los resultados que le sean suministrados por expertos.

3. Separar de una manera eficaz los diferentes papeles del experto encargado de evaluar el peso de la evidencia y del juez o los miembros del jurado.

El trabajo del forense debe orientarse a comparar y evaluar la relación entre los rasgos del habla propios del agresor, o locutor fuente de las grabaciones dubitadas y del sospechoso, o locutor de identidad indubitada del que se obtienen muestras de referencia, y sus conclusiones deben orientarse a evaluar la posibilidad de que ambas identidades se correspondan con el mismo individuo, relacionada con la hipótesis de la fiscalía (θ_P), frente a que se ambos sean individuos diferentes, condición generalmente implícita a la hipótesis de la defensa(θ_D) [5], presentadas originalmente en [35].

Generalmente, el experto forense no está al corriente del resto de pruebas que se presentan en el proceso judicial (como por ejemplo que el sospechoso fuera visto en el lugar del crimen en ese momento por un testigo, o que tuviera una mala relación personal, deudas etc. con la víctima). Esta información (I), que no es suministrada al experto forense, genera unas probabilidades *a priori* de cada una de las hipótesis, y su relación son las denominadas *odds*[1] o razones de probabilidades.

$$O(\theta_P|I) = \frac{P(\theta_P|I)}{P(\theta_D|I)} = \frac{P(\theta_P|I)}{1 - P(\theta_P|I)} \tag{2.1}$$

Sin embargo la información de interés a la hora de sacar conclusiones es la probabilidad aislada de ocurrencia *a posteriori* de cada una de las hipótesis, teniendo en cuenta la información disponible (I) y las observaciones del experto forense respecto de la evidencia(E).

$$O(\theta_P|E, I) = \frac{P(\theta_P|E, I)}{P(\theta_D|E, I)} \tag{2.2}$$

Se debe por tanto, buscar una forma de expresar el peso de la evidencia que no emita conclusiones finales, sino que convierta estas probabilidades *a priori* en probabilidades *a posteriori*, contemplando la nueva evidencia. Una forma de satisfacer estas premisas es hacer uso de una interpretación bayesiana y expresar resultados en forma de *Relación de Verosimilitud* [36, 31], propuestos por primera vez en [37].

[1]El término anglosajón *odds* presenta una traducción al castellano discutida; se ha traducido también como disparidad, razón de posibilidades, razón de oportunidades, oportunidad, razón de momios, etc.

2.5.3. Relación de verosimilitud (LR)

Las conclusiones del experto forense generalmente expresan la probabilidad de la ocurrencia de la evidencia (E) en caso de que el sospechoso sea el locutor de la misma ($P(E|\theta_P, I)$) frente la probabilidad de ocurrencia en caso de que no lo sea ($P(E|\theta_D, I)$).

El teorema de Bayes (Ecuación 2.3) ayuda a relacionar esto con las probabilidades a priori, de manera que sean modificadas por la evidencia para obtener probabilidades a posteriori que la contemplen:

$$O(\theta_P|E, I) = \frac{P(E|\theta_P, I)}{P(E|\theta_D, I)} \times O(\theta_P|I) \tag{2.3}$$

La relación de verosimilitud (Ecuación 2.4) o LR (Likelihood Ratio) mide el peso de la evidencia, y resume de forma eficiente las conclusiones extraídas de la comparación por el experto en una relación de probabilidades:

$$LR = \frac{P(E|\theta_P, I)}{P(E|\theta_D, I)} \tag{2.4}$$

Por tanto, aunque el análisis forense apoye una de las hipótesis, nunca podrá concluir en una afirmación categórica que implicase que el locutor desconocido *es* o *no es* el sospechoso, ni de que lo sea o no *con una cierta probabilidad*, sino que la razón de probabilidad a posteriori de la hipótesis θ_P o θ_D que contemplaban la información disponible I, será X veces mayor o menor que la razón a priori, al tener en cuenta la nueva evidencia E (Ecuación 2.3).

2.5.4. Relación de verosimilitud Multi-Variada (MVLR)

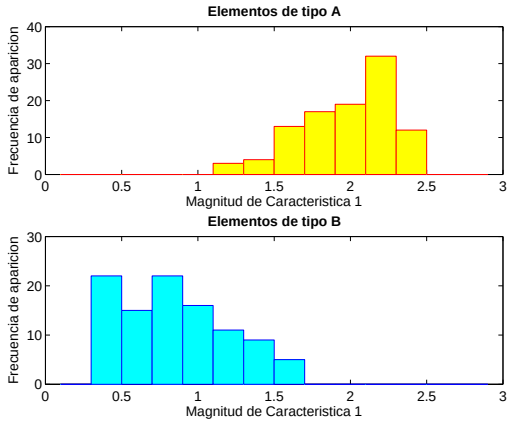
A la hora de valorar la evidencia forense, el análisis de una única característica puede resultar suficiente, por ejemplo porque presente un alto poder discriminativo por sí sola. En este supuesto, la caracterización de la muestra puede reducirse a la medida de una única característica y resumirla en un valor escalar x.

Sin embargo cuando el poder discriminativo de una única característica no es suficiente, o existen otras fuentes de información complementarias, se puede recurrir al uso de varias características a la vez [38]. En estos casos una muestra particular quedaría caracterizada en forma de vector de características $\vec{x}$, compuesto por las medidas de cada característica analizada.

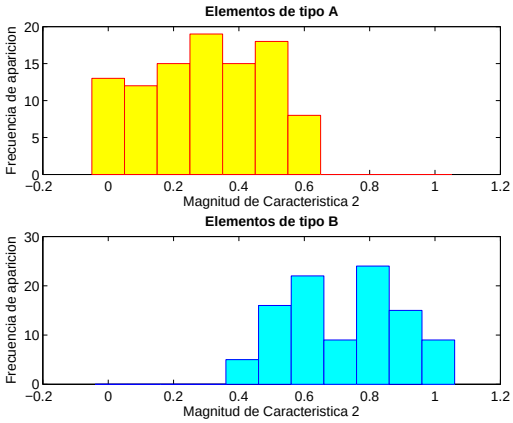
Cuando el uso de un determinado conjunto de características no satisface nuestras expectativas o necesidades de rendimiento, se pueden añadir características nuevas con la intención de mejorarlo. Sin embargo esto no siempre será así. La adición de características altamente correladas con alguna ya analizada, podrá mejorar muy poco o nada el rendimiento. También es posible que la característica nueva sea poco discriminativa y deteriore el rendimiento del sistema. Al margen del rendimiento hay más aspectos que pueden echar atrás a la hora de añadir nuevas características como el coste o la complejidad de su extracción etc.

Sin embargo en ocasiones la adición de una nueva característica puede suponer una gran mejora en el rendimiento de un sistema, en particular cuanto más independiente sea de las características ya contempladas, pues la información nueva será menos redundante.

Por ejemplo, supongamos que se nos facilitan los valores para una característica de un conjunto limitado de elementos que nos disponemos a evaluar. La distribución de dichos valores viene dada por el histograma de la Figura 2.1(a). Cada elemento debe clasificarse como perteneciente al tipo A o al tipo B, de propiedades diferenciadas. En la Figura 2.1(a)



(a) Histograma de valores para la característica 1.



(b) Histograma de valores para la característica 2.

Figura 2.1: Histograma de dos características.

se diferencian claramente tanto los espacios que ocupan los valores de la característica 1 para cada uno de los dos tipos posibles, como una zona de solape que inducirá a errores de clasificación.

Ahora se nos facilita también el conjunto valores de una segunda característica del mismo conjunto de elementos relacionados con los valores de la primera, de tal modo que para cada muestra conocemos ambos valores. El histograma de los valores presentados por las muestras para la característica 2 aislada se puede observar en la Figura 2.1(b).

Se observa que existe también una zona de solape, que se traducirá en ciertos errores al clasificar las muestras utilizando la característica 2 individualmente. Sin embargo si representamos las parejas de valores para cada muestra en un espacio bidimensional, donde cada dimensión represente una característica, obtendremos algo similar a la Figura 2.2.

Analizando las Figuras 2.1(a), 2.1(b) y 2.2, se observa claramente que ambas características son complementarias, y que su uso conjunto mejorará sensiblemente el poder de discriminación de un sistema basado en ellas, de hecho se podría trazar intuitivamente una frontera en el espacio bidimensional que separara eficazmente las zonas de clasificación, sobre las que podríamos asignar tipo A o B a los elementos que ocupen cada una de ellas.

A la hora de analizar varias características discriminativas de manera conjunta, se pueden asumir dos estrategias definidas:

1. Analizar cada una de ellas de manera individual, obteniendo para cada una de ellas un LR tal como se describió en la Sección 2.5.3, y combinarlos, multiplicándolos si están en unidades naturales, o sumándolos en unidades logarítmicas, para obtener un LR general que englobe la información aportada por cada una. Este planteamiento asume independencia estadística entre el conjunto de características analizado, situación que no se da en todos los casos.

2. Un análisis generalizado que aplique una estrategia de análisis multivariado de datos [39, 16], es decir, que analice los datos por agrupación (análisis por clústers, análisis de regresión etc.). Generalmente se obtienen mejores resultados que analizando las características de forma individual.

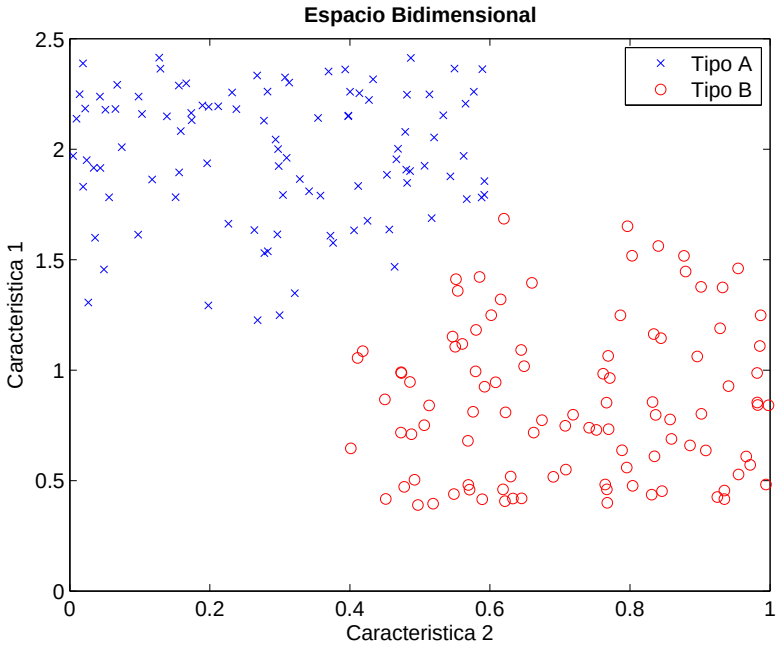


Figura 2.2: Representación bidimensional evaluando conjuntamente dos características.

Un ejemplo de este segundo tipo de estrategias es la fórmula de Aitken & Lucy [16], utilizando estimaciones de características según densidades kernel multivariadas. Existen más trabajos al respecto como por ejemplo [40]y utiliza esta información para descomponer el problema global en varios subproblemas de menor dimensión.

2.6. Evaluación del rendimiento de métodos de reconocimiento forense de locutor

A la hora de evaluar un sistema de reconocimiento, se hacen con él una serie de comparaciones sobre muestras o grupos de muestras de origen conocido. Dichas comparaciones son de dos tipos:

1. **Comparaciones target:** evalúan casos para las que θ_P es cierta a través de comparaciones entre muestras que tienen el mismo origen. Una comparación target debería generar como resultado un LR positivo en unidades logarítmicas, o mayor que la unidad en unidades naturales.

2. **Comparaciones non-target:** evalúan casos para las que θ_D es cierta a través de comparaciones entre muestras que tienen orígenes distintos. Una comparación non-target debería generar como resultado un LR negativo en unidades logarítmicas, o menor que la unidad en unidades naturales.

Una forma inmediata de evaluar el rendimiento del sistema, es contabilizar el numero de FA (falsas aceptaciones, es decir, dar como target una comparación non-target) y FR (falsos rechazos, es decir, dar como non-target una comparación target). Dichos valores dependerán no solo del sistema, sino de un umbral establecido que separe resultados que se clasificarán como target de los que se clasificarán como non-target, por lo que una pareja de tasas de FA y FR sólo determinarán el funcionamiento del sistema en un punto

particular, determinado por el umbral escogido, llamamos discriminación a la separación entre comparaciones target y non-target. Menor discriminación implica una reducción en la tasa de FA para un FR constante o una reducción en la tasa de FR para un FA constante.

Para cierto valor del umbral, las tasas de falsa aceptación y falso rechazo toman el mismo valor, que se denomina EER (*Equal Error Rate*) y que caracteriza el funcionamiento del sistema de forma resumida en un único valor, aunque sólo para un punto de funcionamiento. Sin embargo puede que los objetivos del sistema (como por ejemplo reducir específicamente la tasa de FA o de FR) o un funcionamiento significativamente mejor del mismo en alguna región específica, nos lleven a trabajar con un valor diferente de umbral. A continuación se presentan algunas representaciones que definen el rendimiento del sistema para cualquier punto de trabajo, y no solo basándonos en la discriminación del mismo.

2.6.1. Gráficas DET

Las gráficas DET (*Detection Error Trade-Off*) resumen la discriminación del conjunto experimental de valores de LR en una única curva [41], por lo que la representación conjunta de las curvas de varios sistemas en una única gráfica resulta muy intuitiva, y muy útil para comparar el rendimiento de ambos en cualquier punto de trabajo.

Para elaborar una gráfica DET se evalúa el conjunto de LRs generados por el sistema para comparaciones target y non-target, a lo largo de un conjunto de valores de umbrales, y se calculan las tasas de falsa aceptación y falso rechazo como medida del rendimiento del sistema para cada punto de trabajo establecido por el valor del umbral que separa las puntuaciones que serán evaluadas como target (θ_P es cierta) de las que serán evaluadas como non-target (θ_D es cierta). A continuación, se representan en una gráfica la tasa de falso rechazo frente a la de falsa aceptación para todos los puntos de funcionamiento del sistema. En el origen de coordenadas se establece el valor 0 para ambas. Obtendremos una representación similar a la de la Figura 2.3. En una gráfica DET se puede ver de manera inmediata:

1. El valor de EER (*Equal Error Rate*). Se correspondería con el punto de corte de la curva DET del sistema con una línea diagonal imaginaria que recorrería la gráfica desde el origen de coordenadas y que representaría la ecuación $FA = FR$, definición del valor EER.

2. El rendimiento del sistema en cualquier punto de trabajo. Eligiendo un valor de FA o FR en los ejes y buscando su proyección en la recta, el otro valor es inmediato, al proyectar de nuevo el punto correspondiente de la recta sobre el otro eje. Además es fácil ver para que zonas trabaja mejor en líneas generales: las zonas de mayor rendimiento tienen mayor cadencia hacia el origen de coordenadas.

3. La comparación entre varios sistemas. Al representar juntas 2 o más curvas DET, la proximidad al origen de coordenadas diferenciará el rendimiento de ambos (tanto general como para cualquier zona específica) de manera bastante gráfica. Los puntos de corte entre ellas se corresponden con puntos de trabajo para los que ambos sistemas ofrecen igual rendimiento.

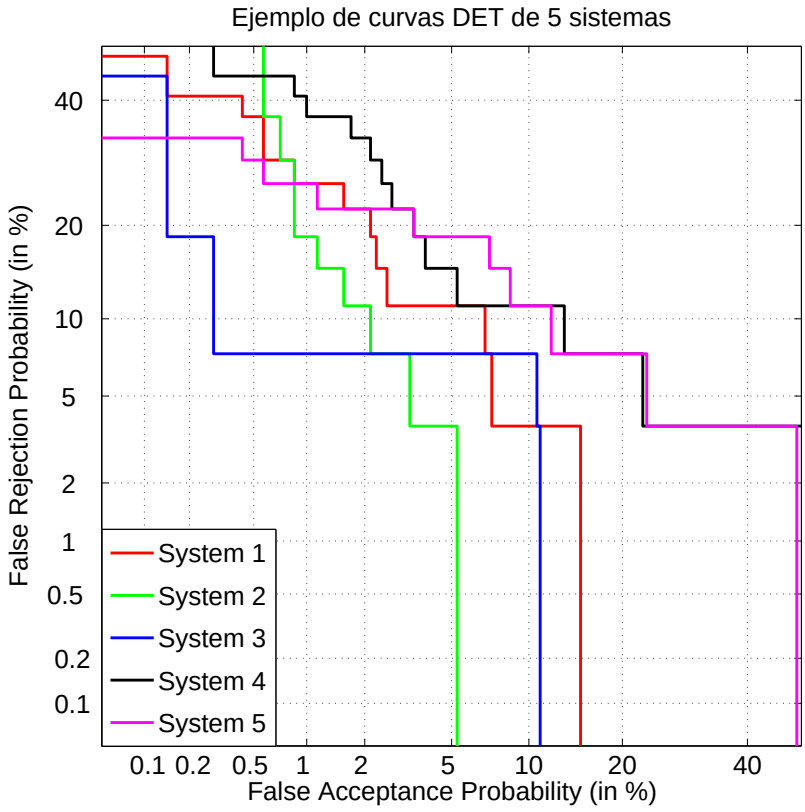


Figura 2.3: Ejemplo de curvas DET para 5 sistemas ficticios.

2.6.2. C_{llr} y C_{llr}^{min}

Para definir el rendimiento de los LR finales generados por el sistema se pueden utilizar valores C_{llr} propuestos en [42]:

$$C_{llr} = \frac{1}{2 \cdot N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{LR_i}\right) + \frac{1}{2 \cdot N_d} \sum_{j_d} \log_2 \left(1 + LR_j\right) \tag{2.5}$$

siendo N_p y N_d el numero de comparaciones (valores LR) para las que θ_P y θ_D (definidas en 2.5.2) son ciertas respectivamente en el conjunto experimental, y que representan por tanto comparaciones target (entre muestras del mismo individuo) y non-target (entre muestras correspondientes a individuos diferentes).

Se puede observar en la Ecuación 2.5 que el C_{llr} es una medida de rendimiento sobre un conjunto de LRs. Cuanto mayor sea C_{llr} peor es el rendimiento del sistema que generó dichos valores.

La pérdida general de rendimiento reflejada por C_{llr} se puede descomponer en:

1. Una pérdida debida a la limitación del poder discriminativo del conjunto experimental sobre el que se trabajaba, en base a la distribución de LRs target y non-target en el espacio de los números reales. Este valor se denomina C_{llr}^{min} y es capaz de resumir una curva DET (Sección 2.6.1) en un único valor, cuanto menor sea éste, mayor poder discriminativo del conjunto experimental (ver Figura 2.4). C_{llr}^{min} representa la máxima optimización alcanzable de C_{llr} (sin alterar el poder discriminativo del conjunto experimental) por el sistema optimizado con un algoritmo PAV (*Pool Adjacent Violators*, para más información ver [43, 44]). Ver [42] para más detalles.

2. Una pérdida por calibración, debida a la posición en la recta real de las relaciones de verosimilitud target y non-target con respecto de los rangos que idealmente ocuparían, si se van a utilizar como grado de apoyo a las hipótesis . La pérdida por calibración puede calcularse según la relación: $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$. Ver [42] para más detalles.

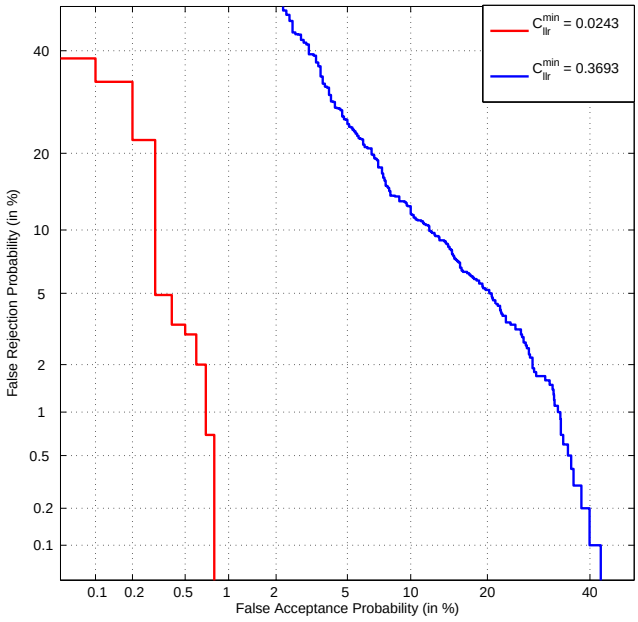


Figura 2.4: Ejemplo de curvas DET y valores de C_{llr}^{min} para los mismos sets de valores.

2.6.3. Gráficas APE

Las gráficas APE (*Applied Probability of Error*) propuestas en [42] son una representación gráfica de rendimiento de un sistema. Son una manera de representar la probabilidad media de error si se escogen umbrales óptimos (Sección 2.6.2). La probabilidad media de error específico se define en la Ecuación 2.6:

$$P_e = P(\theta_P)P_{FR}(-\lambda) + P(\theta_D)P_{FA}(-\lambda), P_1 = logit^{-1}(\lambda) \tag{2.6}$$

siendo P_{FR} y P_{FA} las probabilidades de FA y FR en el valor negativo del Umbral de Bayes λ expresado como:

$$\lambda = log\frac{P(\theta_D)}{P(\theta_P)} \tag{2.7}$$

En una gráfica APE, como en la de ejemplo de la Figura 2.5 se observan claramente los siguientes elementos:

- La línea sólida en color rojo, es la tasa de error real del sistema tal y como está calibrado. Esta es la medida de rendimiento válida. Se puede demostrar ([42]) que el área de esta curva a lo largo de todo su dominio se corresponde con el valor de C_{llr}, y está representado en la altura total de la barra en el diagrama de barras bajo las APEs.

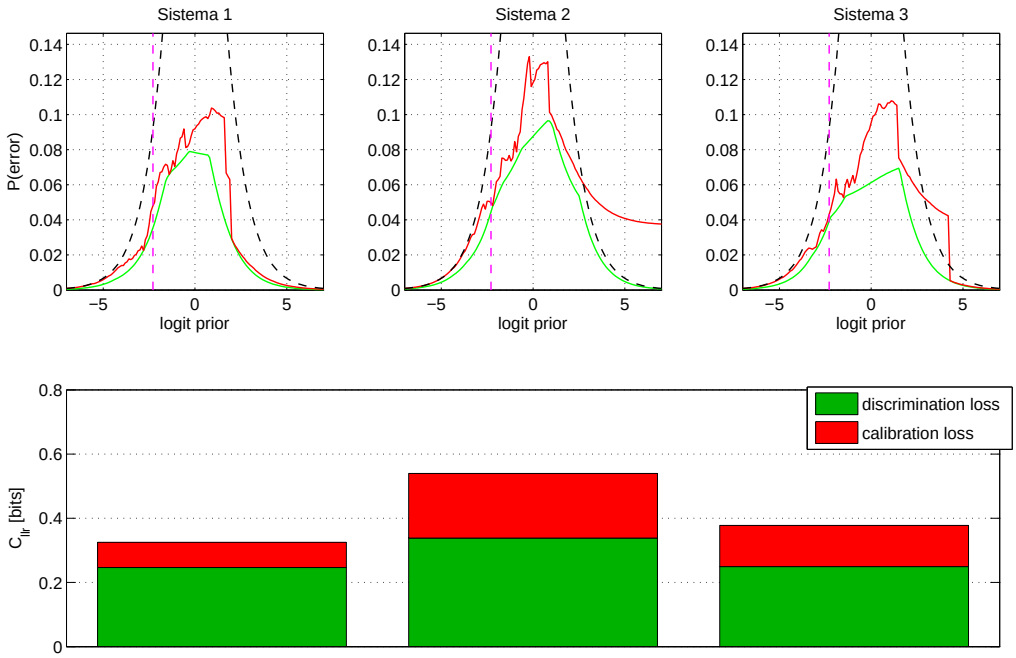


Figura 2.5: Ejemplo de representación APE para 3 sistemas ficticios.

- La línea sólida verde, es la tasa de error optimizada con un algoritmo PAV (*Pool Adjacent Violators* [43, 44]), que representa el sistema de referencia, supuesta una calibración perfecta. Representa la probabilidad de error mínima alcanzable manteniendo la discriminación, supuesto un sistema con calibración perfecta. Se puede demostrar ([42]) que el área de esta curva a lo largo de todo su dominio se corresponde con el valor de C_{llr}^{min}, que está representado en la altura de la sección verde de la barra en el diagrama de barras bajo las APEs.

- La línea negra punteada representa la tasa de error un sistema que no procesa la entrada sino que asigna a la muestra el valor más probable de entrada, y que por lo tanto tiene una probabilidad de error $P_e = min(P(\theta_D), P(\theta_P))$. El área de esta curva es la unidad y no se representa en el diagrama de barras.

- La sección roja del diagrama de barras representa C_{llr}^{cal}, proporcional al área que queda entre las líneas que representan la tasa de error real y la tasa de error optimizada simulando calibración perfecta, o lo que es lo mismo la diferencia entre C_{llr} (altura total de la barra) y C_{llr}^{min} (altura de la sección en verde).

A la hora de analizar una representación APE, conviene recordar que las probabilidades de error disminuyen hasta ser despreciables o incluso 0 al alejarse del origen de abscisas ($P(\theta_D) = P(\theta_P)$, es decir target y non-target son equiprobables), ya que el conocimiento a priori amortigua sensiblemente el error.

2.6.4. Tippett plot

Las gráficas Tippett (ejemplo en la Figura 2.6) propuestas en [45] son otra forma de representar el funcionamiento del sistema. Consisten en la representación de las distribuciones acumulativas en dos curvas diferentes de los logaritmos de los LR generados

por los grupos de comparaciones de las que se sabe a priori que son target y non-target respectivamente.

Existen dos variantes de las gráficas Tippett:

1. Tanto para las comparaciones target como non-target se representa el porcentaje de realizaciones para las que el sistema genera un resultado en forma de log-LR mayor que su proyección en el eje de ordenadas, por lo que se generan dos curvas monótonas decrecientes, que irán de $y = 100\,\%$ para $x = -\infty$ hasta $y = 0\,\%$ para $x = +\infty$.

2. En la que las comparaciones target, para las que valores altos son más deseables, se representa el porcentaje de realizaciones para las que el sistema genera un resultado en forma de log-LR menor que su proyección en el eje de ordenadas, por lo que la curva generada es monótona creciente en este caso pues la acumulación de comparaciones non-target se presenta hacia la izquierda (valores negativos), y la acumulación de comparaciones target hacia la derecha (valores positivos).

   El resultado es el mismo que si se representase la tasa de falsa aceptación $P(FA)$ y la tasa de falso rechazo $P(FR)$ con respecto a un umbral definido por el valor en el eje de ordenadas.

Por tanto, en las gráficas Tippett del segundo tipo, para un punto de la recta, el eje de ordenadas y representa la tasa de comparaciones que se encuentran por encima (para comparaciones target) o por debajo (para comparaciones non-target) del valor de Log-LR definido por el eje de abscisas x. En este tipo de gráficas se puede ver de forma inmediata:

- El EER (*Equal Error Rate*), localizando el cruce de ambas curvas y haciendo su proyección sobre el eje y, que reflejará el punto del trabajo para que las tasas de FA y FR son iguales entre sí, y a su vez al EER. Además, es fácil evaluar el rendimiento en cualquier otro punto de trabajo, pues el movimiento por el eje de ordenadas equivale al desplazamiento del umbral λ del sistema, y las curvas en ese punto representarán la tasa de FA y FR respectivamente.

- También es posible para un resultado (Log-LR) concreto conocer inmediatamente la probabilidad de error de la calificación asignada. Basta con buscar la puntuación en el eje de ordenadas y observando las curvas se ve rápidamente para ese LR la proporción de errores FR o FA (según se haya optado por θ_D o θ_P).

- La distribución que toman los resultados de comparaciones target y non-target. Teniendo en cuenta que cuanto mayor sea el log-LR asignado a una comparación target (y menor a una non-target) mejor será el rendimiento del sistema, será fácil hacerse una idea de cómo éste responde en general para cada tipo de comparaciones, valorando la distribución de la fuerza con la que se apoyan las hipótesis.

- El poder de discriminación del sistema. Por ejemplo en un sistema que consiga separación total, ambas curvas no se cruzarán nunca, y cuanto más separadas estén las regiones que ocupa cada una, mayor será el poder discriminativo. Por el contrario si se cruzan, existe una región de incertidumbre correspondiente con la zona de solape.

- La fuerza con la que se puede llegar a apoyar una hipótesis errónea. Cuanto más se introduzcan una en otra, quiere decir que comparaciones non-target generaron puntuaciones más altas (o comparaciones target puntuaciones más bajas), y por tanto más fuerza pueden llegar a adquirir las clasificaciones incorrectas generadas por el sistema.

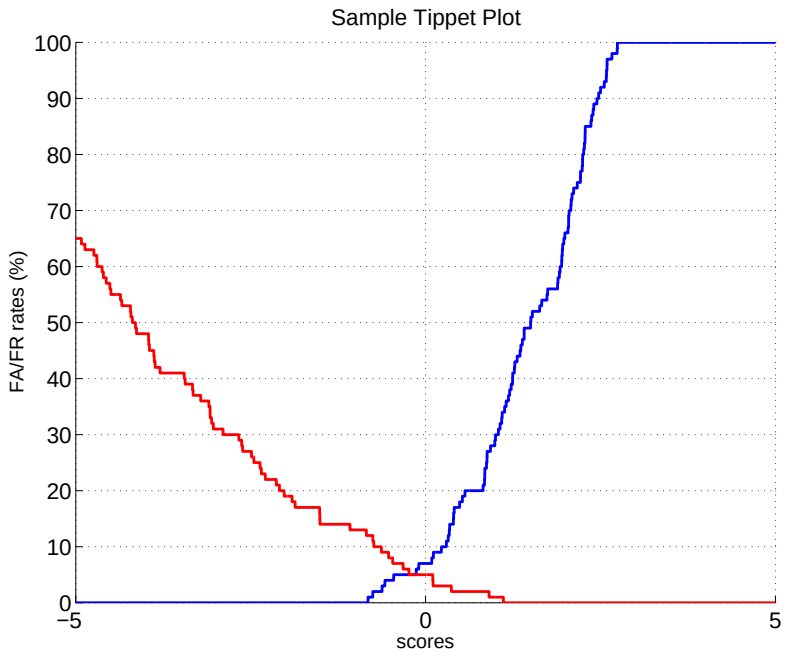


Figura 2.6: Ejemplo de curva Tippett para un sistema ficticio.

3

Seguimiento automático de formantes y selección de características para reconocimiento forense de locutor

3.1. Introducción al Capítulo

En este Capítulo se describe el diseño de un sistema completo de reconocimiento forense de locutor con extracción automática de formantes. Se empleará un seguidor automático de formantes para la extracción de la trayectoria en las regiones de interés de los 3 primeros formantes F1, F2 y F3. A continuación se ajustan curvas expresables paramétricamente a las trayectorias de estos formantes, quedando descritos por un conjunto reducido de valores o características que los describen de forma bastante completa, y que serán tomados por el sistema como elementos discriminantes del locutor.

También se describe una estrategia de selección de características, que partiendo de las anteriores, trata de analizar individualmente cada una de las características y elegir exclusivamente aquellas que ofrezcan un mayor rendimiento con el objetivo de mejorar el sistema. Los resultados de la elección de características también pueden ser interpretados de manera fonético-acústica, para evaluar de forma sencilla las propiedades de diptongos y formantes que conllevan un mayor poder de discriminación.

La contribución de la primera estrategia es la automatización de la supervisión a cargo de un experto en la fase de extracción de formantes frente a propuestas semi-automáticas o de extracción manual. La selección de características ofrece una mejora en el rendimiento de los sistemas, y hace factible un análisis fonético-acústico de la aportación de cada característica.

Los algoritmos desarrollados han sido publicados en la 10^{th} Annual Conference of the International Speech Communication Association (ISCA): Interspeech 2009 Brighton (http://www.interspeech2009.org), congreso internacional de máximo impacto en el área de tratamiento de voz, en el artículo [1] reproducido en el Apéndice A:
A. de Castro, D. Ramos, and J. Gonzalez-Rodriguez, *"Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking"* in *Proc. of Interspeech*, 2009.

3.2. Contribución: automatización del seguimiento de formantes

Para la extracción automática de formantes de las zonas señaladas que contienen las unidades de habla de interés particular (diptongos o vocales), se ha empleado la herramienta seguidora de formantes descrita en [10] para MATLAB^{TM}. La Figura 3.1 muestra un ejemplo del uso de esta técnica.

La herramienta de seguimiento de formantes está basada en una estimación de las frecuencias formantes por medio de un proceso de Gauss-Markov, modelado según:

$$x_{t+1} = Fx_t + w_t \tag{3.1}$$

siendo $x_t = (f_1, ..., f_K, b_1, ..., b_K)^T$ un vector de estado, resultado de parametrizar la envolvente espectral en cada ventana temporal t, y que contiene las frecuencias y los anchos de banda de los K primeros formantes. F es una matriz de transición entre estados y w_t es una secuencia de ruido blanco. Para esa región temporal, los primeros N coeficientes cepstrales LPC (*Linear Predictive Coefficients*) son observados, y relacionados con el vector x_t según la Ecuación 3.2. En este caso se emplearon 12 coeficientes cepstrales y ventanas de 20ms, ambos valores son ajustables en la herramienta. De acuerdo con [10], la relación entre el enésimo coeficiente cepstral $y_t[n]$ y el vector de estado x_t viene dado por:

$$y_t[n] = \frac{1}{N}\sum_{k=1}^{N} exp\left(\frac{\pi n}{f_s}x_t\left[k+K\right]\right) cos\left(\frac{2\pi n}{f_s}x_t\left[k\right]\right) \tag{3.2}$$

donde f_s es la frecuencia de muestro de la señal. La especificación del modelo se completa con la función no lineal $h(\cdot)$ que según se define en la Ecuación 3.2 calcula $y_t = (y[1], ..., y[N])^T$ (coeficientes cepstrales LPC) a partir de x_t (vector de frecuencias y anchos de banda de formantes).

$$y_t = h(x_t) + v_t \tag{3.3}$$

siendo v_t una secuencia de ruido blanco. Se asume que ambas funciones de ruido son incorreladas de tal forma que satisfagan $E(v_i w_j^T) = 0$ para todo i y j. Finalmente la distribución $p(x_t|y_{1:t})$ se calcula para las frecuencias formantes en t, condicionado a la información de la señal previamente observada. Su media se utiliza como estimación y la varianza para definir la incertidumbre de los valores.

Para obtener detalles sobre el funcionamiento de esta herramienta o en su desarrollo, consultar [10]

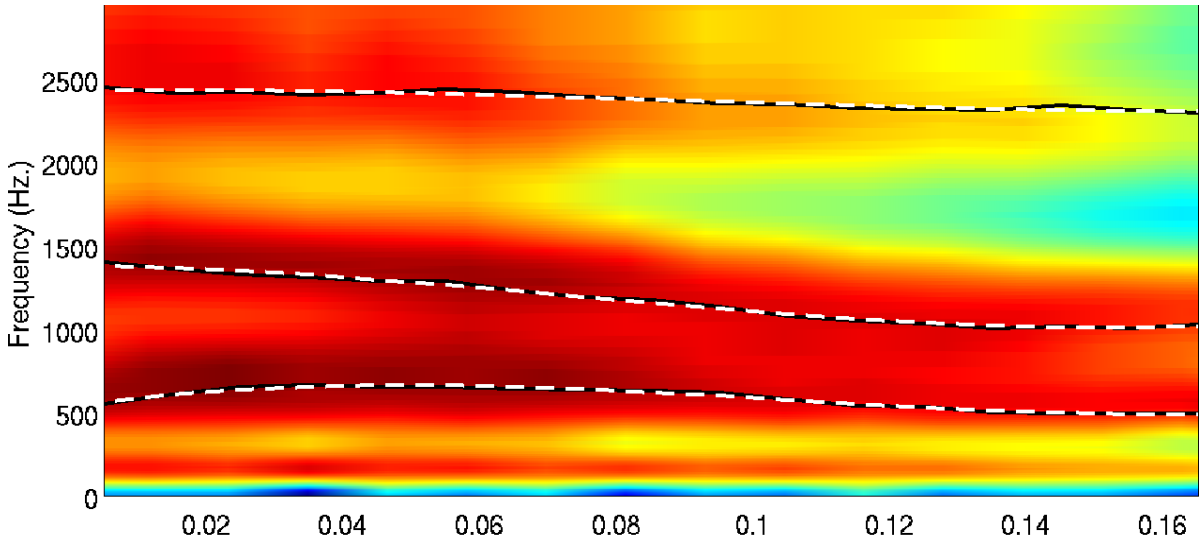


Figura 3.1: Ejemplo de extracción automática de formantes y ajuste a curva paramétrica. El diptongo bajo análisis es /oʊ/. La curva paramétrica a la que se ajusta está descrita por una ecuación polinómica de grado 3. Las líneas sólidas son los formantes extraídos, y las líneas punteadas son las reconstrucciones a partir de los parámetros de ajuste.

3.3. Métodos de ajuste paramétrico

El seguidor automático de formantes nos devuelve la trayectoria extraída automáticamente para cada formante muestreada de acuerdo a la duración del etiquetado manual previa del diptongo y al ancho de ventana especificado, de tal forma que para cada diptongo analizado se generan 3 vectores de datos correspondientes a los 3 primeros formantes, y otro vector de igual número de elementos que los anteriores, que contiene los valores de tiempo correspondientes, y que para eliminar dependencia temporal en el ajuste paramétrico se desplaza en todo caso para que tenga valor inicial 0.

Para determinar la naturaleza del ajuste paramétrico se emplearon dos enfoques. Cada una de las alternativas contempladas a la hora de afrontar el ajuste paramétrico materializan estrategias radicalmente diferentes de representación paramétrica de curvas:

- **Ajuste polinómico:** la trayectoria de cada formante del diptongo bajo análisis, se ajusta a una curva descrita por una ecuación polinómica que describe la frecuencia en función del tiempo. Partiendo de la expresión paramétrica general de una curva, se ajusta el valor de los coeficientes de tal modo que la diferencia con la trayectoria extraída sea mínima. Se contemplan polinomios de grados 3 (Ecuación 3.4a) y 2 (Ecuación 3.4b), que presentan 4 y 3 coeficientes respectivamente teniendo en cuenta el término independiente, y que pueden resultar óptimas alternativamente, acorde al número de puntos críticos esperados para la trayectoria genérica de cada diptongo. Nos referiremos a estas estrategias como *poly*2 y *poly*3 respectivamente.

$$\mathbf{a}x^3 + \mathbf{b}x^2 + \mathbf{c}x + \mathbf{d} = 0 \rightarrow (a, b, c, d) \tag{3.4a}$$

$$\mathbf{a}x^2 + \mathbf{b}x + \mathbf{c} = 0 \rightarrow (a, b, c) \tag{3.4b}$$

- **Transformada Discreta del Coseno (DCT):** Al vector de valores se le aplica una transformada DCT unidimensional [46], y lo convierte en otro vector de valores correspondientes a los coeficientes que definen dicha transformación (Ecuación 3.5). Los primeros valores del vector transformado son los de mayor relevancia a la hora de reconstruir la señal, por lo que se puede reducir el peso de la información sin pérdidas sensibles al reconstruir (la codificación de imágenes JPEG está basada en esta cualidad de la transformación DCT). El elemento 0 define el offset de la trayectoria, el elemento 1 define la amplitud del componente equivalente a medio ciclo de un coseno, el elemento 2 define la amplitud del componente equivalente a un ciclo completo de coseno, el elemento 3 define la amplitud del componente equivalente a ciclo y medio de coseno, y así sucesivamente. Para este experimento tomaremos, en dos alternativas, los elementos 0, 1, y 2 en la estrategia denominada *DCT*2 y los elementos 0, 1, 2 y 3 en la estrategia denominada *DCT*3.

$$\mathbf{X_c}(k) = \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos(\frac{k2\pi n}{N}) \rightarrow \begin{cases} (X_c(0), X_c(1), X_c(2), X_c(3)) \\ (X_c(0), X_c(1), X_c(2)) \end{cases} \tag{3.5}$$

Adicionalmente se contemplan diferentes alternativas de procesado previo de los datos de los que se dispone. El objetivo del procesado previo es adecuar (y validar la adecuación) de la información disponible para optimizar los resultados, empleando técnicas evaluadas en experimentos anteriores [9] como susceptibles de aportar una mejora funcional al sistema. Las estrategias de procesado previo contempladas son las que se relatan a continuación:

- **Duración ecualizada o natural:** Ecualizar la duración de los diptongos puede suponer una mejora, ya que para el mismo diptongo pronunciado por el mismo usuario, la diferencia de duración entre una realización y otra (que puede llegar a ser bastante significativa) puede desvirtuar la información intrínseca a la curva. Para contrarrestar este efecto, todas las trayectorias de formantes bajo análisis se interpolan linealmente de manera que resulten en vectores de 126 valores, correspondiente a una escala temporal unificada de 0 a 250ms muestreada cada 2ms.

En el caso de emplear una escala de tiempos natural, se utiliza directamente la información extraída, de tal forma que el eje temporal vaya de 0ms a la duración directa del diptongo en la muestra, y el vector de trayectoria será de cualquier longitud, acorde a la duración y el tiempo de ventana.

- **Frecuencia en escala logarítmica o natural:** Variar la escala, para que deje de ser lineal, puede suponer también una mejora en el modelo como se evaluó en [9]. Para ello se aplica logaritmo neperiano a los valores nominales de frecuencia extraídos, y que definen la trayectoria del formante, resultando en un vector de igual longitud de valores en log-Hertzio representativos del formante, que en adelante, y para esta alternativa (frecuencia en escala logarítmica) será tratado de igual manera que el vector original, que se empleará en la alternativa con escala de frecuencias natural.

- **Descarte de F1:** El habla telefónica, limita el espectro que puede ocupar la voz a la región comprendida entre 350 y 3500Hz aproximadamente. El límite inferior estimativo de la región principal formántica es de 200Hz para varias vocales, por lo que en sistemas de análisis de habla telefónica se tiende a descartarlo directamente al estar desvirtuado y ofrecer un rendimiento pobre, no obstante no impide la inteligibilidad de la conversación aunque sí afecta a la percepción del habla. En principio el descarte sistemático de F1 implicaría una simulación más realista del sistema.

Para evaluar la mejora o no implicada por cada una de las citadas estrategias, se ha simulado el sistema para cada diptongo, con todas las combinaciones de estrategias alternativas de ajuste paramétrico y tratamiento previo de señal, y posteriormente se siguieron 3 criterios distintos de selección descritos en la Sección 3.4.2 de acuerdo con el rendimiento mostrado por cada conjunto de condiciones.

3.4. Esquema de ajuste paramétrico utilizando seguimiento automático de formantes

En esta sección se describe de forma global un sistema forense de reconocimiento de locutor basado en ajuste paramétrico que aplica las mismas técnicas utilizadas en [9] con la salvedad de que en lugar de extracción semi-automática de formantes supervisada por experto, se utiliza seguimiento automático de formantes, tal y como se puede ver en la Figura 3.2, con el objetivo de comparar resultados para poder valorar el rendimiento de la extracción automática de las trayectorias de los formantes.

3.4.1. Obtención de relaciones de verosimilitud

En este PFC se ha utilizado la implementación en MATLAB^{TM} [47] disponible en http://geoff-morrison.net/ de la fórmula de Aitken & Lucy [16], para obtener relaciones de verosimilitud utilizando estimaciones de características según densidades kernel multivariadas. Se evalúa la diferencia entre muestras dubitadas e indubitadas con respecto a la repetitividad de características en una distribución estimada sobre una población tipo apropiada. La variación entre muestras del mismo tipo es estimada a través de una distribución normal, y entre muestras de diferente tipo es estimada a través de un modelo kernel de densidad de probabilidad de acuerdo al siguiente modelo:

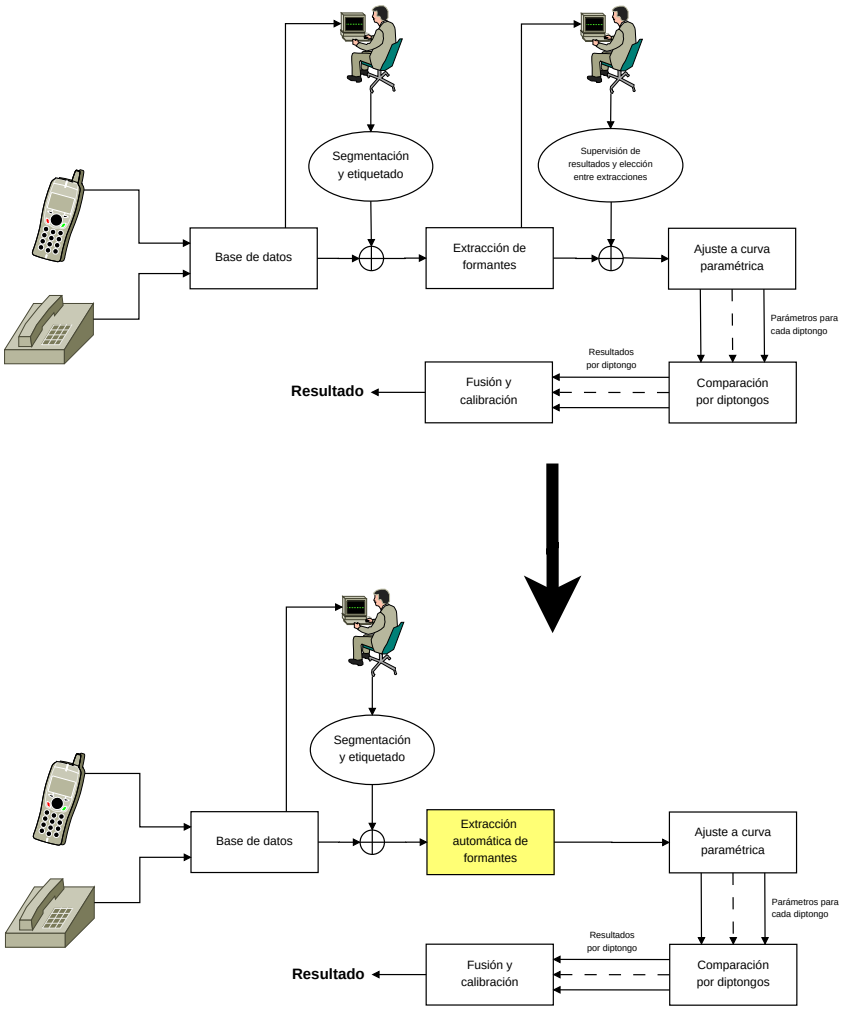


Figura 3.2: Esquema de aplicación de extracción automática de formantes sobre el sistema de ajuste paramétrico.

$$LR = \frac{p(\vec{x},\vec{y}|\,\theta_P,I)}{p(\vec{x},\vec{y}|\,\theta_D,I)} = \frac{\int p(\vec{x},\vec{y}|\,\mu,\theta_P,I)\,p(\mu|\,\theta_P,I)\,d\mu}{\int p(\vec{x}|\,\mu,\theta_D,I)\,p(\mu|\,\theta_D,I)\,d\mu \int p(\vec{y}|\,\mu,\theta_D,I)\,p(\mu|\,\theta_D,I)\,d\mu} \tag{3.6}$$

donde θ_P es la hipótesis de la fiscalía (*El sospechoso es fuente de las grabaciones dubitadas*), θ_D es la hipótesis de la defensa (*Otro individuo de la población tipo relevante es fuente de las grabaciones dubitadas*), $\vec{x}$ e $\vec{y}$ son vectores de características a comparar del material de habla dubitada y de control respectivamente, I es el resto de información disponible, y μ es una variable de integración. Para más detalles del modelo consultar [36].

3.4.2. Criterios de extracción de características

En la Sección 3.3 se definieron diferentes condiciones de tratamiento previo de señal y de naturaleza de ajuste paramétrico, a saber, la duración del diptongo puede haber sido ecualizada o no, se puede haber aplicado logaritmo neperiano a los valores de frecuencia de la trayectoria o haber hecho uso de los valores naturales, se puede haber empleado F1 F2 y F3 o tan sólo F2 y F3, y los valores finales pueden ser fruto de haber ajustado

la curva a una expresión polinómica de grado 2 o grado 3, o bien haber aplicado la transformada discreta del coseno a la trayectoria y haber escogido desde el coeficiente 0 (offset) hasta el 2^o o 3^{er} coeficiente.

De cara a seleccionar un conjunto adecuado de características, se ha realizado una valoración preliminar sobre cada diptongo. Para cada comparación se ha generado un valor LR por cada diptongo. Este procedimiento se repite de manera completa para cada método descrito en la sección 3.3, obteniendo para cada método y diptongo, un set de resultados completo en forma de matriz cuadrada de LRs de NxN elementos, siendo N el número de usuarios de la base de datos.

Estos resultados servirán para evaluar el poder discriminativo de cada método para cada diptongo bajo estudio. Se toma como medida de eficiencia el C_{llr}^{min} calculado tal y como se describe en la Sección 2.6.2 a partir de los LR multidimensionales sin calibrar.

Con esta valoración preliminar se han desarrollado tres sistemas diferentes de acuerdo a tres estrategias distintas de extracción de características en base a esta medida de rendimiento, a saber:

- **Mejor rendimiento individual (BEST_IND):** de entre todas las posibles combinaciones de ajuste paramétrico y tratamiento previo de señal, se escogerá para cada diptongo, aquella que haya generado mejores resultados para él en particular, resultando en 5 criterios independientes (uno por diptongo) pero no necesariamente diferentes (una combinación particular puede ser la que mejor funciona para varios diptongos). Este criterio de selección debería ser a priori el que mejores resultados finales ofrezca, ya que selecciona los ajustes que proporcionan un mejor rendimiento para cada diptongo.

- **Mejor rendimiento colectivo (BEST_ALL):** se elegirá una única combinación de condiciones para todos los diptongos, que será la que mejor C_{llr}^{min} haya generado en media. La elección de ser de este criterio no busca mejorar los resultados, sino evaluar la pérdida de rendimiento al generalizar el sistema al uso de un único conjunto de características que funcione aceptablemente bien para todos los diptongos, en busca de un sistema más sencillo y más generalizable cuyo rendimiento no se aleje demasiado del sistema BEST_IND, correspondiente a la elección específica del mejor conjunto para cada diptongo.

- **Criterio para extracción semi-automática (HUMAN_AUTO):** corresponde al conjunto de características seleccionadas por el experto humano en [9], escogiendo las condiciones que generaron un mejor c_{llr}^{min} individual para cada diptongo empleando trayectorias de formantes extraídas de forma semi-automática, tal como se describe en la Sección 2.3.1. El objetivo del sistema implementado siguiendo este criterio, es efectuar una comparación directa con el experimento llevado a cabo en [9] sobre la misma base de datos y un sistema similar pero aplicando extracción semi-automática de formantes supervisada por experto. De tal forma se puede analizar la bondad del extractor automático de formantes y evaluar el coste en forma de deterioro del rendimiento de una mayor automatización.

3.4.3. Fusión y calibración de resultados

Tras la obtención de valores de LR para cada diptongo resulta necesario efectuar una fusión de tal forma que los diferentes resultados individuales puedan resultar en uno solo que los resuma de forma eficaz. También es necesario un proceso de calibración [48], que ajuste los valores de los LR y establezca el grado correcto de apoyo a una hipótesis de

acuerdo con el funcionamiento del sistema, compensando el grado de desviación de los resultados generados por el mismo. La calibración se realiza apoyándose en comparaciones entre muestras de control conociendo a priori que θ_P o θ_D son ciertas respectivamente, y compensa las desviaciones entre muestras que provienen de diferentes diptongos y que no son interpretables probabilísticamente. El resultado final perseguido es un único valor de LR o Log-LR, que exprese el peso de la evidencia, arrojado por el sistema.

Se siguen tres estrategias para la fusión y calibración de los diferentes resultados, de los que se dispone para cada comparación, en uno solo, que represente lo mejor posible los resultados que para ella ofrece el sistema. Se han seguido tres estrategias diferentes:

- **Fusión lineal con regresión logística** tal y como se describe en [42]. Con esta estrategia se fusionan los LR obtenidos para cada diptongo, y se calibra el resultado general. La principal ventaja es que los procesos de fusión y calibración de resultados se realizan en un único paso, sin embargo el número de muestras de referencia necesarias es elevado y crece exponencialmente con la dimensionalidad de la fusión, es decir, con el número de resultados a fusionar.

- **Suma de Log-LR y calibración a posteriori**. En primer lugar se suman todos los log-LR calculados para cada diptongo obteniendo un único valor para cada comparación. A continuación, se aplica una etapa posterior de calibración para compensar la falsa presunción de independencia entre muestras.

- **Calibración a priori y suma de Log-LR calibrados**. Cada conjunto de resultados obtenido para cada diptongo, se calibra usando regresión logística [42], de tal modo que para una comparación se tiene un Log-LR calibrado por cada diptongo, que se sumarán en un único valor que represente la comparación, y que no será calibrado de nuevo, por ser resultado de la suma de valores calibrados individualmente.

En algunos casos, lo reducido del número de muestras disponibles imposibilitó la ejecución de la fusión lineal con regresión logística para el conjunto total de diptongos, por lo que se aplicó sobre combinaciones seleccionadas de diptongos. Se intentó adaptar las demás estrategias para obtener resultados comparables a estos, ya que considerando el sistema extrapolable a bases de datos mayores, se evitaría esta situación y dicha estrategia sí sería finalmente útil para obtener resultados válidos.

Para todas las calibraciones se empleó una técnica Jackknife [49], consistente en que para cada comparación, se omite el uso en todos los procesos de entrenamiento de todas aquellas muestras disponibles del sujeto o los sujetos bajo comparación como muestras de referencia. Con ello se consigue que, a la hora de efectuar la comparación, sus muestras sean tratadas como datos nuevos, evitando el uso de muestras de los individuos que se van a comparar en la elaboración del modelo y obteniendo resultados coherentes con identidades dubitadas y, por tanto, más realistas.

3.5. Contribución: selección de características

Una contribución de este proyecto consiste en un procedimiento de selección de características, que utiliza los ajustes de trayectorias de formantes a curvas paramétricas, pero no de manera global, como se ha hecho en la Sección 2.3, sino escogiendo aquellas características (definidas como los parámetros que describen la curva que mejor se adapta a la trayectoria de cada formante) que mejor rendimiento individual ofrecen. Este procedimiento de selección de características reemplaza la fase de fusión, al generar por sí mismo un único resultado a partir de resultados individuales.

Este sistema persigue la mejora general del rendimiento, que sobre el conjunto de entrenamiento será sensible, al escoger las características que ofrecen mayor poder discriminativo para este conjunto en particular, según el esquema de la Figura 3.3. Sin embargo puede ser extrapolado probando el conjunto de características obtenidos sobre bases de datos diferentes o sobre muestras que no se incluyeran en el conjunto de entrenamiento.

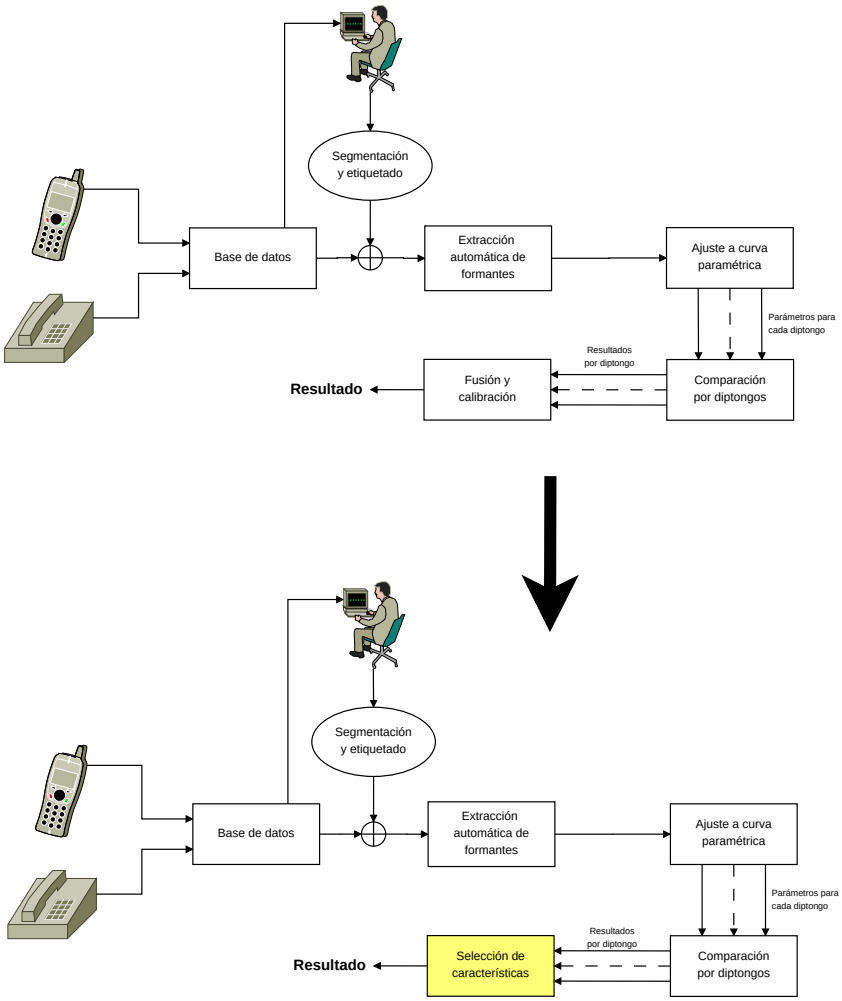


Figura 3.3: Esquema de integración de selección de características sobre el sistema con ajuste paramétrico y extracción automática de formantes.

Otro objetivo del sistema puede ser la obtención de nueva información sobre el poder de discriminación de ciertas características en particular, lo que puede tener una clara interpretación fonético-acústica. Es decir, una representación gráfica de las características seleccionadas, puede servir para averiguar de manera sencilla, qué diptongo, qué formante, o qué característica, diferencia mejor un locutor de otro, o específicamente dentro de cualquier diptongo en particular, por ejemplo, que formantes y/o características son más discriminativos, y estudiar su relación con las trayectorias genéricas de los formantes para ese diptongo, etc. Esto es muy útil para el experto porque le permite comprobar de forma sencilla qué características son más discriminativas, y esta información puede ser utilizada para el desarrollo o mejora de sistemas de reconocimiento forense de locutor.

Para este sistema los valores obtenidos para los diferentes criterios de selección descritos en 3.4.2 también se replican, de cara a comparar los resultados obtenidos siguiendo el esquema de selección de características con resultados obtenidos con el esquema de ajuste paramétrico con seguimiento automático de formantes descrito en la Sección 3.4.

3.5.1. Procedimiento de selección de características

Partimos por tanto, para cada criterio de extracción, de una serie de características (valores paramétricos) que definen las curvas que, para cada diptongo, siguen los formantes en cada aparición del diptongo bajo estudio pronunciado por un locutor en particular. Tendremos por tanto un vector de características fruto de dicho ajuste.

Como novedad, cada una de las características del vector va a ser tratada de manera individual y no agrupadas de tal forma que se obtenga un único resultado por diptongo.

De tal forma, se efectuaron comparaciones entre usuarios, usando cada característica individualmente, obteniendo relaciones de verosimilitud unidimensionales con cada característica. Así, para cada característica y diptongo, se dispone ahora de un conjunto de $N \times N$ resultados (N es el número de usuarios de la base de datos), sobre el que se calcula el rendimiento sin considerar calibración (al no estar calibrado) por medio de un valor C_{llr}^{min} (descrito en la Sección 2.6.2).

Ahora se elabora una tabla con todos los valores C_{llr}^{min} para tener un fácil acceso a ellos. En la tabla, cada fila es uno de los diptongos bajo estudio, y cada columna representa una característica, agrupadas de forma conjunta según el formante al que pertenecen para una fácil interpretación. La Tabla 3.1 muestra un ejemplo ficticio de tabla.

	F1			F2			F3		
Diptongo	C1	C2	C3	C1	C2	C3	C1	C2	C3
Diptongo 1	0.8943	0.8313	0.6391	0.7847	0.8073	0.786	**0.6271**	0.6863	0.7142
Diptongo 2	0.8742	0.819	0.7831	0.8436	0.8776	0.7412	0.6351	0.631	0.7069
Diptongo 3	0.9288	0.8934	0.7866	0.7955	0.7725	0.7077	0.6383	0.769	0.642
Diptongo 4	0.7279	0.664	0.6391	0.8919	0.7173	0.7248	0.6927	0.7306	0.6698
Diptongo 5	0.8979	0.8138	0.7323	0.8586	0.8966	0.8659	0.8204	0.7606	0.7241

Tabla 3.1: *Ejemplo de tabla de C_{llr}^{min} calculados individualmente para 5 diptongos y 9 características (3 frecuencias formánticas con 3 características cada una) para utilizar en la selección de características. En negrita el valor mínimo, que se escogería en primer lugar.*

Llegados a este punto, teniendo almacenados el conjunto de resultados para cada característica y una vez elaborada la tabla con los C_{llr}^{min} que caracteriza el poder de discriminación de cada una de ellas, se procede con el siguiente algoritmo:

1. En primer lugar se selecciona la característica que tenga el C_{llr}^{min} mínimo de la tabla, es decir, aquella que mejor rendimiento individual ha ofrecido en las comparaciones, y se define un sistema básico que sólo hace uso de esta característica. Por lo tanto los resultados temporales del sistema son el conjunto de NxN Log-LRs calculados previamente para esta característica y diptongo. Se marca esta característica como seleccionada y su C_{llr}^{min} se elimina de la tabla.

2. A continuación, se busca la característica con menor C_{llr}^{min} entre las restantes en la tabla, y se prueba si mejora el sistema actual de la siguiente manera: se fusionan los resultados, sencillamente sumando los Log-LR actuales del sistema a los Log-LR individuales de la característica en fase prueba. Se calcula un nuevo C_{llr}^{min} para el conjunto de resultados fusionados con la nueva característica y se compara con el C_{llr}^{min} anterior del sistema. Si mejora (es decir C_{llr}^{min} disminuye), se elige definitivamente y su C_{llr}^{min} se elimina también de la tabla. En caso contrario, se deshace la fusión y se continúa probando la siguiente característica con menor C_{llr}^{min} entre las

aún disponibles y no probadas, y así sucesivamente hasta que se encuentre alguna que mejore el sistema. Cabe destacar, que aunque una característica se pruebe y no sea seleccionada por no mejorar el sistema, no se elimina de forma definitiva, ya que se contempla la posibilidad de que sí mejore el sistema en algún paso posterior.

3. El paso anterior se repite hasta que no queden características para escoger, o el sistema no mejore con ninguna característica de las restantes disponibles. En el último paso se habrá evaluado la adición al sistema de todos los parámetros disponibles, y ninguno de ellos lo ha mejorado.

La evolución del C_{llr}^{min} general del sistema, sigue forzosamente una trayectoria monótona decreciente, al ser el valor de referencia de rendimiento, y necesariamente disminuir en cada paso (en caso contrario el paso se desharía). Se puede ver un ejemplo de la evolución del valor C_{llr}^{min} para un sistema en la Figura 3.4.

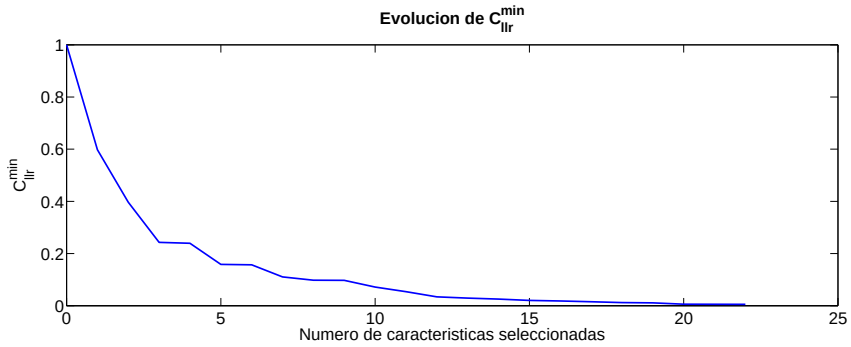


Figura 3.4: Evolución de C_{llr}^{min} global del sistema con la adición progresiva de parámetros siguiendo el esquema descrito de selección de características.

En base a la selección final de características, se puede elaborar una tabla que represente la elección o no de cada una de ellas para facilitar una interpretación fonética de los resultados, como la de la Figura 3.5.

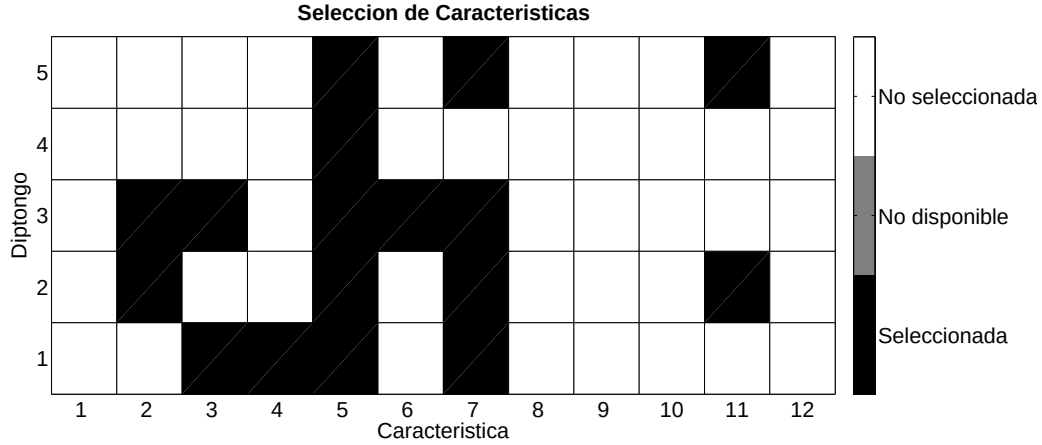


Figura 3.5: Tabla representativa de la elección o no de cada parámetro con el objetivo de facilitar una interpretación fonético-acústica intuitiva.

Siguiendo este procedimiento de selección de características, se podría complementar con una etapa final de post-calibración que compensase la asunción de independencia

estadística entre las diferentes características, lo que no es necesariamente cierto. Sin embargo en este PFC no ha sido considerado como necesario a causa de los bajos errores de calibración en los resultados finales, y de hecho se observó en simulaciones que un proceso de calibración disminuía el rendimiento.

3.5.2. Selección de características mediante Jackknife y extrapolación de resultados

Para tratar de darle generalidad a la aplicación de la selección de características sobre bases de datos pequeñas, y emular el comportamiento que el sistema tendría con muestras nuevas (no empleadas en el entrenamiento), se procede a aplicar un procedimiento Jackknife para efectuar selección de características para, en cada comparación, no tener en cuenta muestras de usuarios implicados en la misma (individuos de la base de datos tomados como agresor y/o sospechoso).

Sin embargo para testear por completo la generalidad, y aprovechando la disponibilidad de una segunda base de datos de diferentes usuarios y obtenida en condiciones distintas, se puede aplicar a esta el conjunto de características seleccionado sobre la primera base de datos, con el objetivo de probar la generalidad de una estrategia de extracción entrenada sobre otra base de datos obtenida en condiciones diferentes, y sobre otros diptongos (o vocales) diferentes. Esto se ha llevado a cabo en este PFC, y sus resultados están contenidos en la sección 4.6.3.

4

Resultados y comparación con procedimientos semi-automáticos

4.1. Introducción al Capítulo

En este Capítulo se explican los resultados de este PFC, tanto finales como intermedios si éstos son considerados de algún interés o tienen cierta relevancia para otros resultados posteriores. Además se detalla el marco experimental, incluyendo tanto las descripciones de las bases de datos empleadas como el protocolo experimental seguido para evaluar los sistemas.

En primera instancia se exponen y analizan los resultados obtenidos con el esquema de ajuste paramétrico con seguimiento automático de formantes descrito en 3.4, de cara a comparar ambos y evaluar el funcionamiento del extractor automático de formantes de [10] descrito en 3.2. Se detallan los resultados parciales considerados relevantes, y los resultados finales, obtenidos en primer lugar sobre la base de datos de *Kinoshita & Osanai 2006* [20], y a continuación aquellos obtenidos sobre la base de datos de *Zhang 2007* [23]. Cuando proceda y sea posible, se comparará con los resultados obtenidos con el proceso de extracción semi-automática de [9] descrito en 2.3.1.

También se exponen y analizan los resultados obtenidos sobre las mismas bases de datos, aplicando el esquema de selección de características, contribución original de este proyecto, descrito en la Sección 3.5. Cuando proceda y sea posible, se comparará con los resultados del esquema de ajuste paramétrico con seguimiento automático de formantes pero sin selección de características descrito en la Sección 3.4 para evaluar el procedimiento de selección de características.

4.2. Marco experimental

Los sistemas basados tanto en esquema de ajuste paramétrico y selección de características descritos en el Capítulo 3 se aplicaron sobre dos bases de datos etiquetadas diferentes para contrastar su funcionamiento. En primer lugar se aplicó sobre la base de datos de *Kinoshita & Osanai 2006*, de locutores masculinos de inglés australiano, y sobre la que se contaba con los resultados del experimento llevado a cabo en [9] para efectuar comparaciones y valoraciones de la extracción automática de formantes.

El sistema desarrollado se aplicó también a la base de datos de *Zhang 2007*, de locutores masculinos de chino mandarín, más extensa en cuanto a número de locutores, y realizada en un ámbito menos controlado (codificación GSM de habla conversacional, frente a grabación microfónica de lectura de frases de control de *Kinoshita & Osanai 2006*).

4.2.1. Base de datos de *Kinoshita & Osanai 2006*

Consiste en grabaciones de audio microfónicas con calidad de laboratorio, extraídas de un corpus de grabaciones previamente descrito en [20] y utilizadas en [9]. Se dispone de locuciones para 27 individuos de género masculino, de entre 19 y 63 años de edad (la mediana del grupo es 39 años). El idioma de las locuciones es inglés australiano, del que todos los individuos son hablantes nativos. Se les hará pronunciar palabras monosilábicas en inglés que contengan uno de los diptongos objeto de estudio, a saber, /aɪ/, /eɪ/, /oʊ/, /aʊ/ y /ɔɪ/. La relación de palabras empleadas puede encontrarse en la Tabla 4.1.

En cada grabación, el locutor pronuncia una frase, en base a una tarjeta que le ha sido enseñada con una palabra. Por ejemplo, cuando al locutor se le enseña una tarjeta con la palabra *BIDE*, el individuo pronunciará la frase: "Bide, B-I-D-E spells bide", es decir, la palabra en cuestión, su deletreo, y de nuevo la misma palabra, por lo que tendremos

	Contexto					
Diptongo	/h_/	/h_t/	/h_d/	/b_/	/b_t/	/b_d/
/aɪ/	high	height	hide	buy	bite	bide
/eɪ/	hay	hate	haydes	bay	bait	spade
/oʊ/	hoe	hote	hoed	bow	boat	bode
/aʊ/	how	-	how'd	bough	bout	bowed
/oɪ/	(coy)	hoytes	-	boy	-	buoyed

Tabla 4.1: *Relación de palabras empleadas para cada uno de los diptongos bajo estudio de la base de datos de* Kinoshita & Osanai 2006.

dos pronunciaciones válidas del diptongo a estudiar, al principio y al final de la frase respectivamente.

Cada usuario fue grabado en dos sesiones separadas aproximadamente dos semanas. En cada sesión el orden en que les fueron mostradas las tarjetas que contenían las palabras fue aleatorio. Para cada palabra se registraron dos realizaciones por sesión. Las grabaciones fueron realizadas con un micrófono SONY^{TM} ECM-MS907 y un grabador Edirol R-1, generando una señal digitalizada a 44.1kHz.

4.2.2. Base de datos de *Zhang 2007*

La descripción de esta base de datos se encuentra en [23]. Los datos se recopilaron de 64 hablantes de chino mandarín, jóvenes y de género masculino. Todos los participantes han crecido en la ciudad de Shenyang, y sus edades están comprendidas entre 19 y 23 años. Todos ellos son estudiantes en la Universidad de Policía Criminal de China (http://www.ccpc.edu.cn/english/).

Aunque los locutores fueron grabados en tres sesiones sin embargo sólo se utilizarán las dos primeras al no estar la tercera completamente etiquetada en el momento de desarrollo del proyecto. La distancia entre las dos primeras sesiones es de 3 semanas aproximadamente.

La base de datos fue recopilada por asistentes de investigación, un par de años mayores que los individuos que la componen. Los asistentes de investigación llamaron por teléfono interno de la universidad a los hablantes y le realizaron una serie de preguntas naturales del tipo "¿Cómo te llamas?", "¿Cuál es tu número de teléfono?", "¿Cuál es la dirección de tu universidad?".

Las grabaciones fueron realizadas con el sistema de teléfono interno de la universidad, empleando teléfonos modelo KCM HCD9999P/TSDL, que incorpora una unidad analógica de grabación en cinta magnetofónica integrada. Las grabaciones fueron entonces digitalizadas utilizando un equipo SANYO^{TM} DC-PT70, y almacenadas como archivos de sonido de 16 bits PCM a una frecuencia de muestreo de 11.025kHz.

Para cada sesión y usuario, se identificaron y etiquetaron diez apariciones acentuadas del fonema /i/ y seis apariciones acentuadas del fonema /y/. El criterio para que fueran acentuadas o no, es que la vocal tuviera una duración superior al umbral de 40ms.

Existen y se encuentran en la base de datos debidamente localizadas y etiquetadas aparición de otras vocales o diptongos, disponibles en las grabaciones realizadas de habla espontánea. Sin embargo en este proyecto no se llegaron a utilizar, replicando el experimento llevado a cabo en [23]. La Tabla 4.2 incluye la relación de todos los diptongos de la base de datos con un número elevado de apariciones, y el número de contextos diferentes en que aparece.

Diptongos	
a (×4)	iiz (×1)
an (×1)	ing (×3)
ao (×1)	iou (×3)
e (×4)	iz (×1)
en (×1)	ou (×1)
er (×1)	u (×4)
i (×4)	uo (×1)
iao (×2)	y (×4)
ie (×2)	

Tabla 4.2: *Relación de diptongos disponibles para estudio de la base de datos de* Zhang 2007 *y entre paréntesis la cantidad de contextos en que aparece cada uno de ellas.*

4.2.3. Protocolo Experimental

Para cada diptongo bajo estudio segmentado por el experto sobre el material disponible (tanto si es una muestra dubitada de cuyo locutor la identidad es desconocida, una muestra indubitada de control del sospechoso o una muestra de un individuo de la población tipo) se genera un vector de características con todos los coeficientes de las ecuaciones paramétricas generados (en adelante *características*). Este vector de características puede tener una longitud diferente para cada diptongo, de acuerdo con el grado del ajuste a curva paramétrica de cada estrategia de extracción. Por tanto para cada individuo y diptongo, tendremos un conjunto de vectores, donde cada vector individual representa una ocurrencia de ese diptongo para la que el individuo en cuestión es el locutor.

Se efectuarán todas las comparaciones posibles entre los individuos de cada base de datos. El conjunto de ocurrencias de cada diptongo en el material dubitado, tomadas todas ellas de una misma sesión, se compara con el conjunto de ocurrencias del mismo diptongo en el material indubitado, tomadas de otra sesión diferente. Por lo tanto, como se dispone de al menos 2 sesiones completas en ambas base de datos, en cada comparación toda la información empleada sobre el agresor provendrá de la primera de ellas, mientras toda la información sobre el sospechoso provendrá de la segunda de ellas. Por tanto sobre cada par de individuos se pueden realizar 2 comparaciones totalmente independientes, en las que cada uno de ellos produzca las tomas dubitadas respectivamente.

Además este planteamiento utiliza para las comparaciones target (el mismo individuo es el autor de las tomas dubitadas e indubitadas) utiliza la misma cantidad de habla e igualmente distanciada temporalmente que para las non-target (entre individuos diferentes), usando el mismo número de muestras en ambos casos, y siempre de una sesión frente a otra. Los resultados serán por tanto representativos, al seguir el mismo procedimiento para ambos tipos de comparaciones (target y non-target), y al emplear sesiones diferentes suficientemente separadas en el tiempo para comparaciones target (en vez de ocurrencias diferentes en la misma sesión, caso en el que no se daría variabilidad inter-sesión). Además cada pareja de individuos diferentes, generará dos resultados de comparaciones non-target, en el primero uno de ellos es el agresor y el otro el sospechoso, y en el segundo se intercambian los roles. Ninguna aparición concreta de diptongo es empleada en ambos procesos, por lo que se podrían considerar como dos casos totalmente independientes al utilizar material completamente distinto.

4.3. Esquema de ajuste paramétrico con seguimiento automático de formantes sobre la base de datos de *Kinoshita & Osanai 2006*

En esta Sección se va a analizar el rendimiento del sistema de ajuste paramétrico descrito en 3.4 que hace uso del extractor automático de formantes de [10] descrito en 3.2 sobre la base de datos de *Kinoshita & Osanai 2006* descrita en la Sección 4.2.1. Este esquema se basa en el ajuste de curvas paramétricas a las trayectorias formánticas y el uso de los parámetros que definen dichas curvas. La base de datos de *Kinoshita & Osanai 2006* es una base de datos de habla microfónica inglesa por hablantes australianos.

El objetivo de esta sección es la evaluación de la degradación del enfoque de ajuste paramétrico por el uso de un extractor automático de formantes.

4.3.1. Rendimiento individual de los diptongos

Una primera medida del rendimiento del extractor automático de formantes se puede tomar a partir de los resultados obtenidos de realizar comparaciones entre individuos utilizando cada diptongo de manera individual.

Esta comparación de los resultados generados por el uso individual de las trayectorias de los formantes extraídas automáticamente para cada diptongo, se puede interpretar como una primera valoración inmediata de la calidad de dichas extracciones, especialmente comparándolo con el rendimiento ofrecido por un sistema conocido que efectúe un procedimiento similar pero basado en extracción semi-automática de calidad contrastada, realizada (o supervisada) por expertos humanos.

En esta Sección se muestran diagramas de barras para cada diptongo de la base de datos de *Kinoshita & Osanai 2006*, que muestran el rendimiento (sin calibrar) en forma de C_{llr}^{min} y C_{llr} ofrecido por cada combinación posible de ajuste polinómico y tratamiento previo de señal.

Para todas las muestras de cada diptongo en particular, los resultados en forma LR se calcularon haciendo una comparación multidimensional de los vectores de parámetros que definen las curvas que se ajustaron a las trayectorias extraídas de los formantes, que para una misma estrategia de ajuste paramétrico siempre serán de la misma longitud.

También se reflejan los resultados de [9] en forma de diagramas de barras de C_{llr}^{min} y C_{llr} calculados para procedimientos similares, pero basadas en trayectorias de formantes obtenidas con el procedimiento semi-automático descrito en 2.3.1.

Pueden servir por tanto para efectuar una valoración de la capacidad discriminatoria de los diptongos usados individualmente. Sobre la misma base de datos, una extracción más precisa ofrece un mayor rendimiento, por lo tanto indirectamente se puede evaluar la bondad del extractor comparando resultados obtenidos con extracción automática de formantes frente a resultados obtenidos con extracción semi-automática de cuya calidad se tiene certeza.

En las gráficas se emplea nomenclatura abreviada para definir cada estrategia, según F1 se mantenga (F1 F2 F3) o se descarte (F2 F3). El ajuste paramétrico puede ser polinómico grado 2 (poly2) o grado 3 (poly3), o una transformación DCT de la que aparte del coeficiente 0 se cogen los 2 (DCT2) o los 3 (DCT) primeros coeficientes. La duración del diptongo puede ser ecualizada (EQ) o natural (noEQ), y la escala de frecuencias puede ser natural (Hz) o logarítmica (logHz).

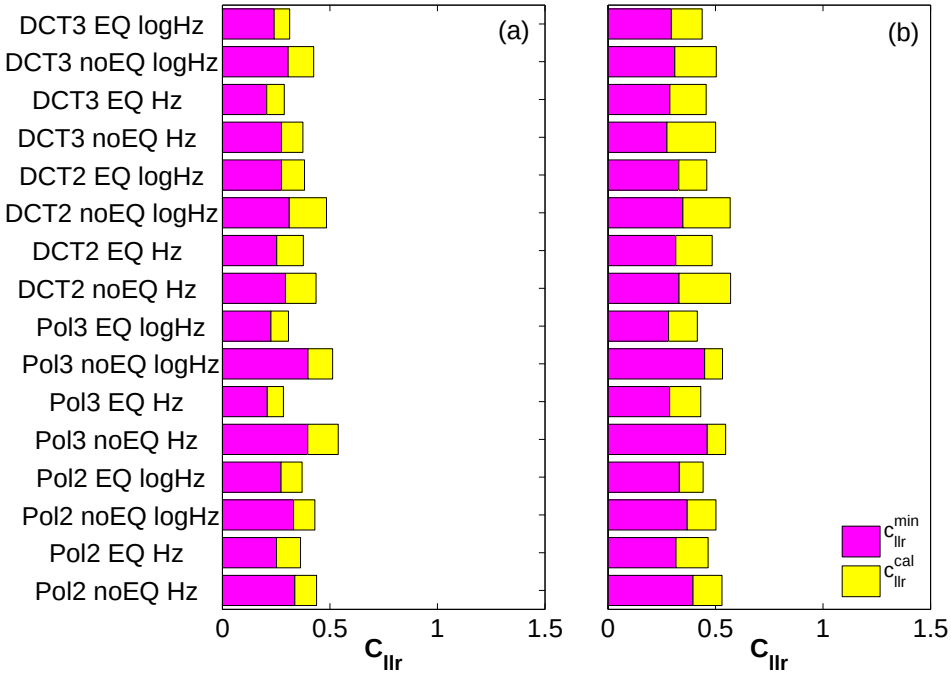


Figura 4.1: Diagrama de barras representativo del rendimiento de las posibles estrategias de extracción para el diptongo /aɪ/ con extracción automática de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

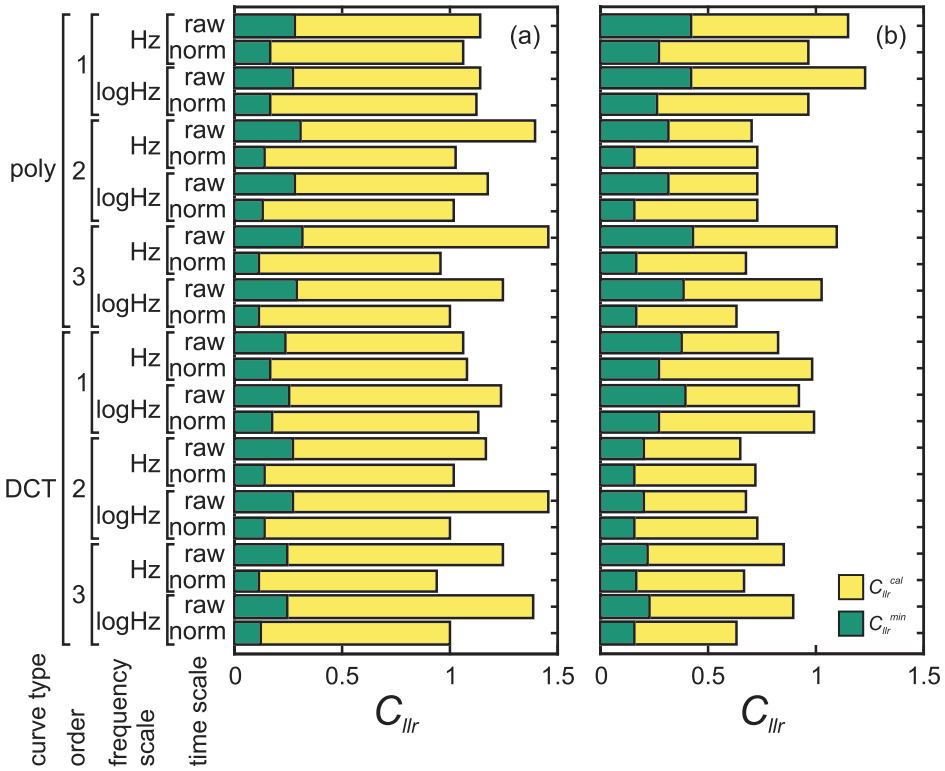


Figura 4.2: Diagrama de barras de [9] representativo del rendimiento de las posibles estrategias de extracción para el diptongo /aɪ/ con extracción semi-automática supervisada de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

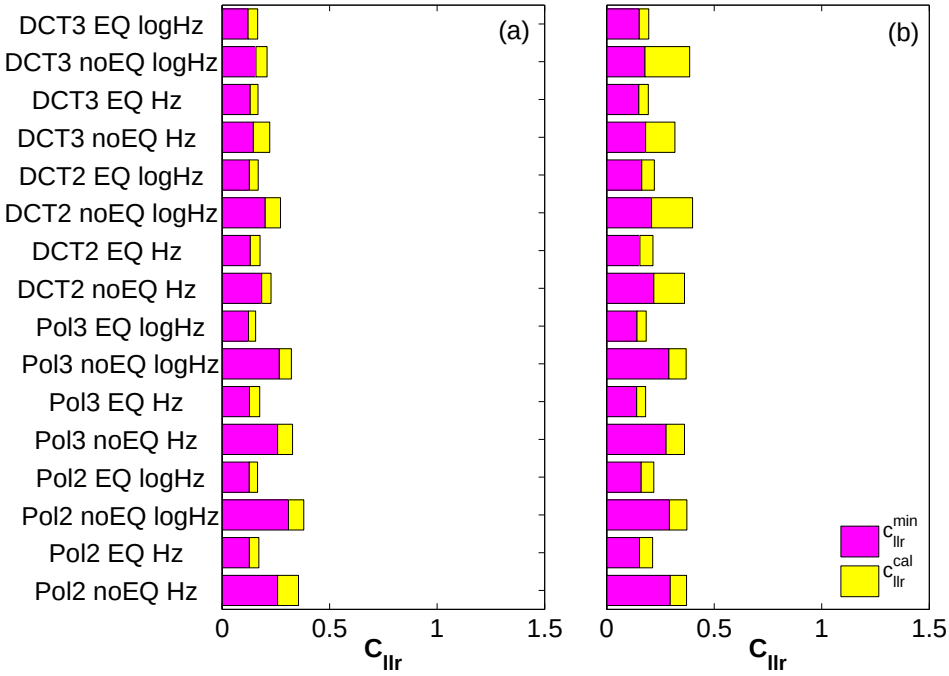


Figura 4.3: Diagrama de barras representativo del rendimiento de las posibles estrategias de extracción para el diptongo /eɪ/ con extracción automática de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

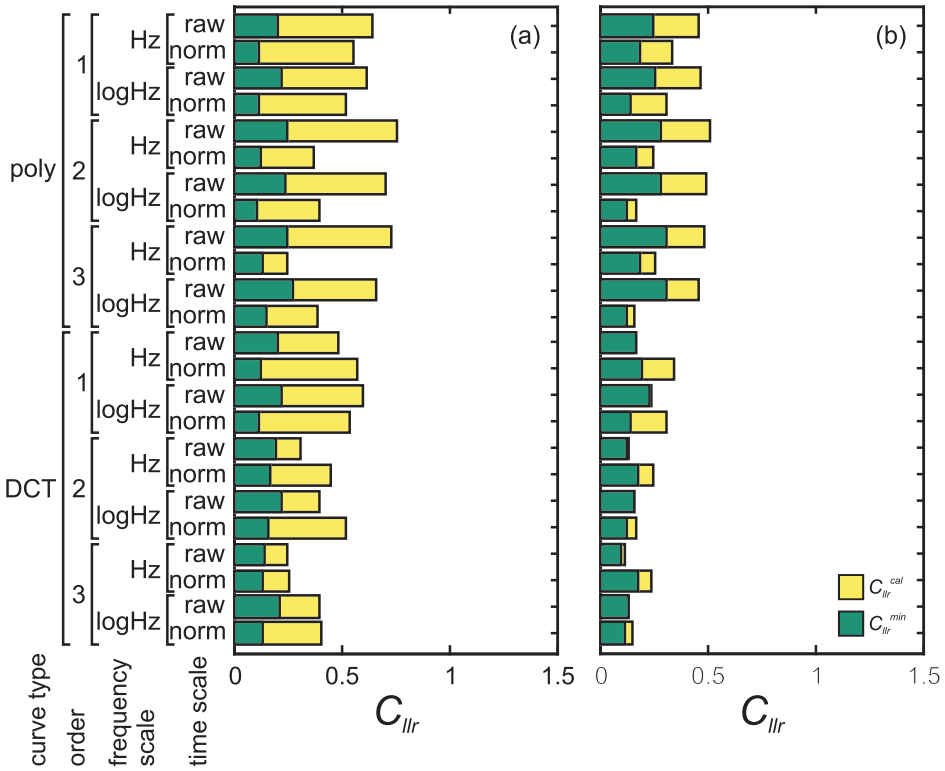


Figura 4.4: Diagrama de barras de [9] representativo del rendimiento de las posibles estrategias de extracción para el diptongo /eɪ/ con extracción semi-automática supervisada de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

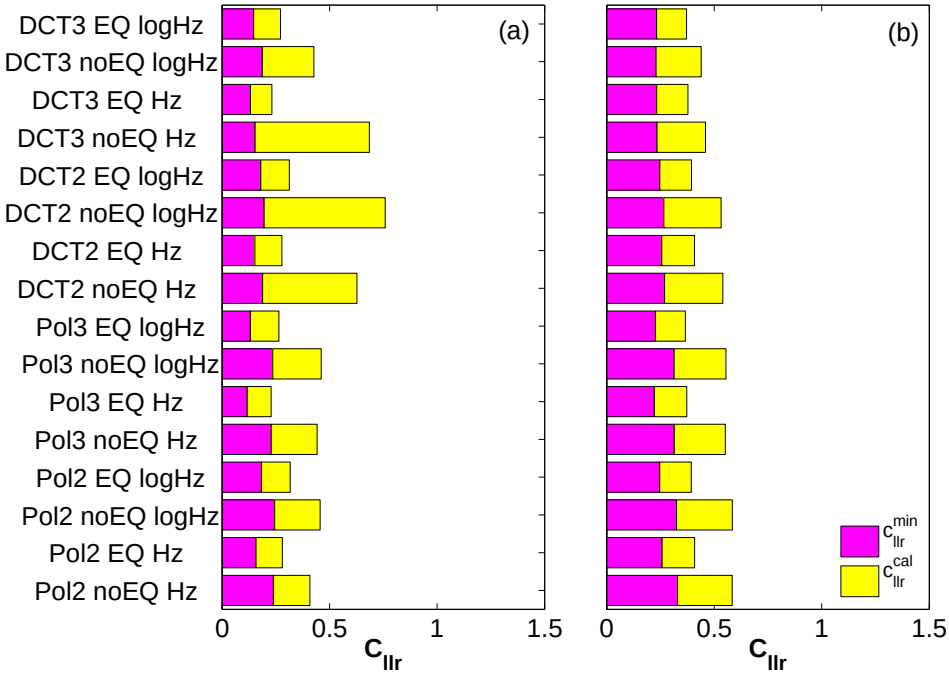


Figura 4.5: Diagrama de barras representativo del rendimiento de las posibles estrategias de extracción para el diptongo /oʊ/ con extracción automática de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

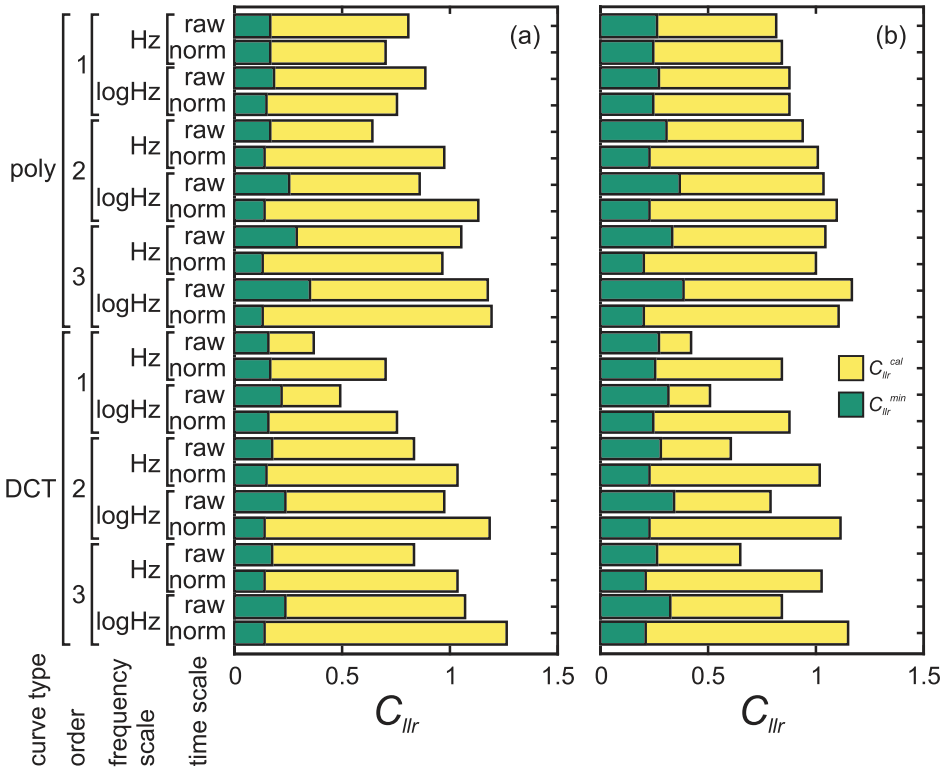


Figura 4.6: Diagrama de barras de [9] representativo del rendimiento de las posibles estrategias de extracción para el diptongo /oʊ/ con extracción semi-automática supervisada de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

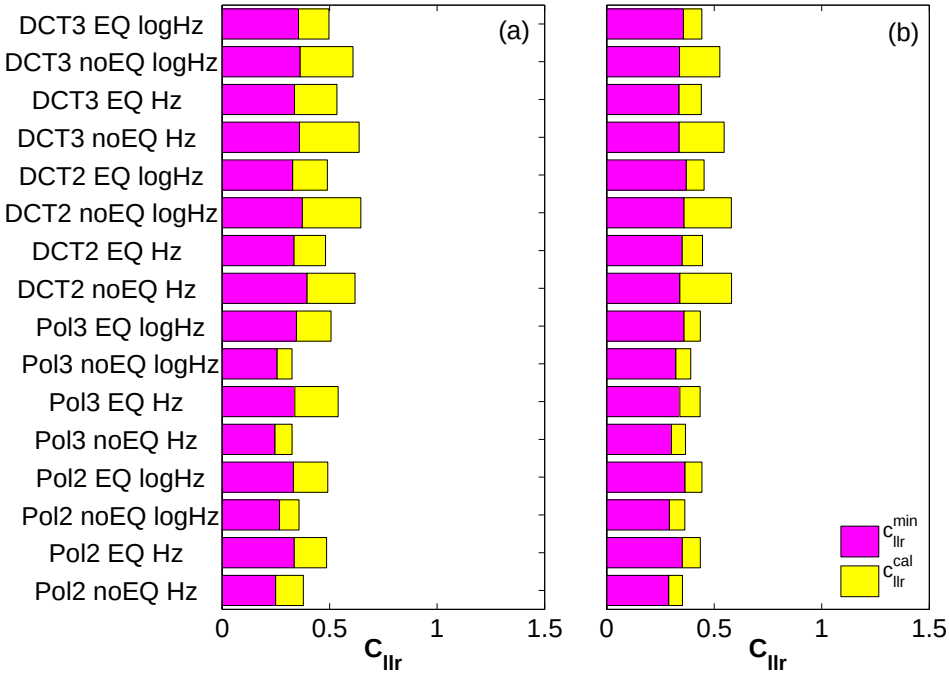


Figura 4.7: Diagrama de barras representativo del rendimiento de las posibles estrategias de extracción para el diptongo /aʊ/ con extracción automática de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

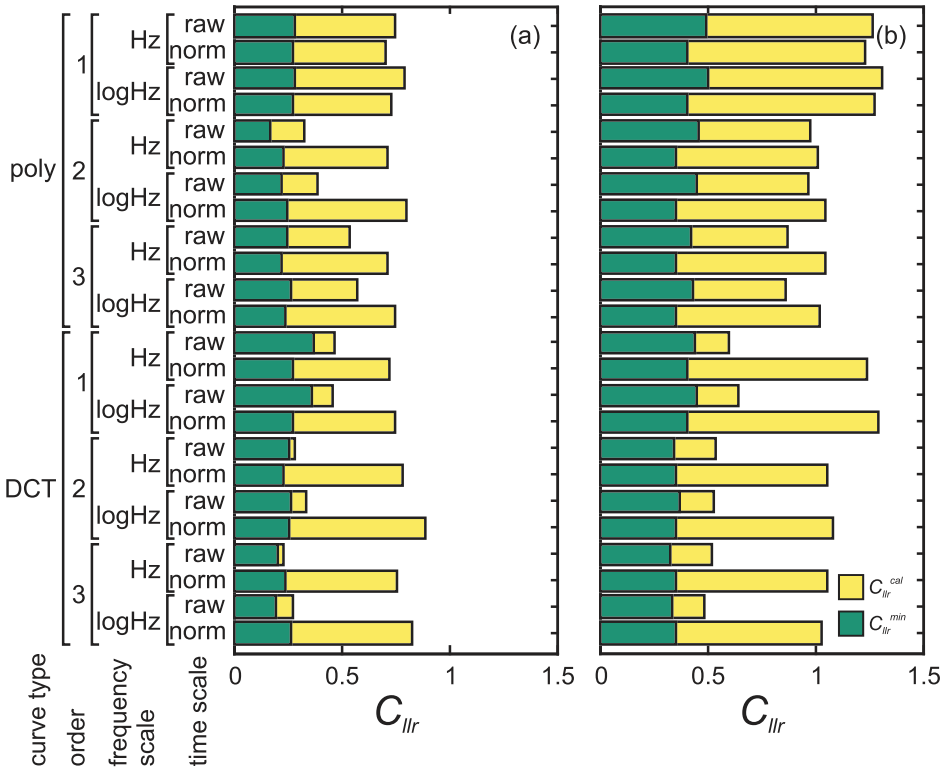


Figura 4.8: Diagrama de barras de [9] representativo del rendimiento de las posibles estrategias de extracción para el diptongo /aʊ/ con extracción semi-automática supervisada de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

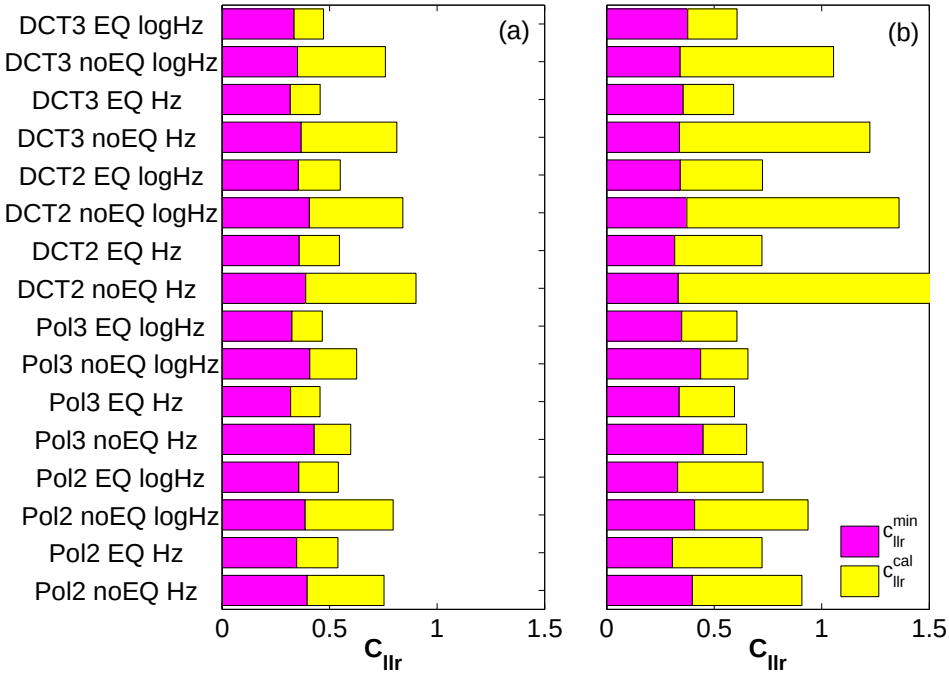


Figura 4.9: Diagrama de barras representativo del rendimiento de las posibles estrategias de extracción para el diptongo /oɪ/ con extracción automática de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

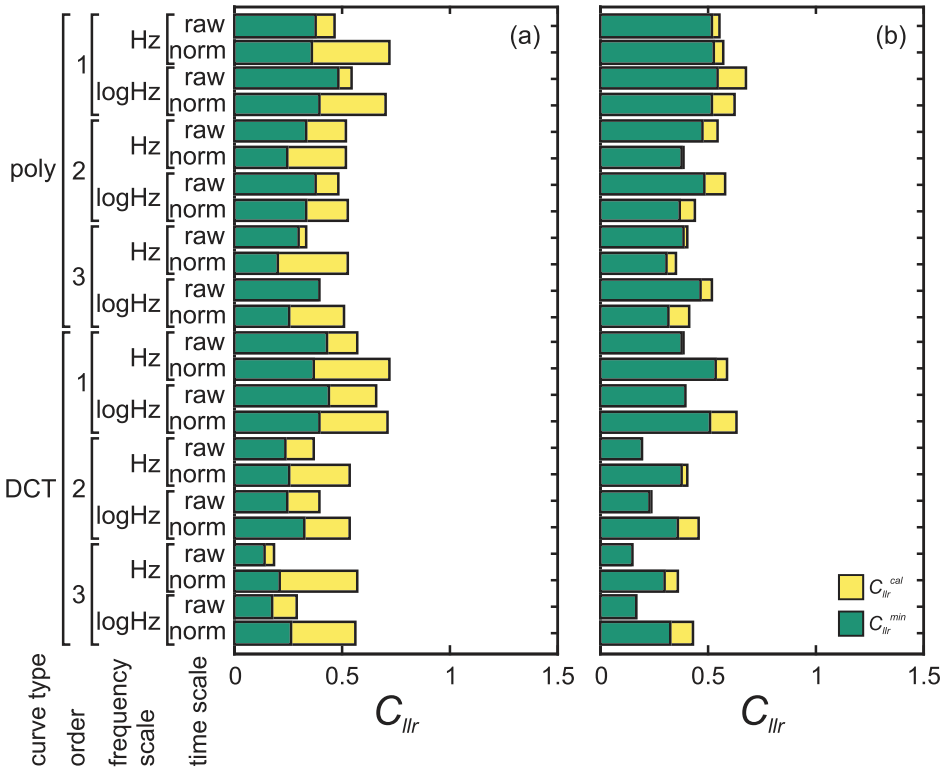


Figura 4.10: Diagrama de barras de [9] representativo del rendimiento de las posibles estrategias de extracción para el diptongo /oɪ/ con extracción semi-automática supervisada de formantes. (a) Usando F1, F2 y F3. (b) Usando F2 y F3.

Se observa que los resultados no están muy distantes de los de [9]. El valor de C_{llr}^{min} se puede comparar directamente ya que no es afectado por procesos de calibración, y mide realmente el poder discriminativo. C_{llr}^{min} toma valor entre 0 (distinción total entre comparaciones target y non-target) y 1 (no se procesan las muestras y devuelve siempre el mismo valor). Se observa que los valores en general están más alejados de 1. Los valores de C_{llr}^{min} no son muy distantes de los de [9] por lo que podemos suponer que el rendimiento del extractor automático es aceptable.

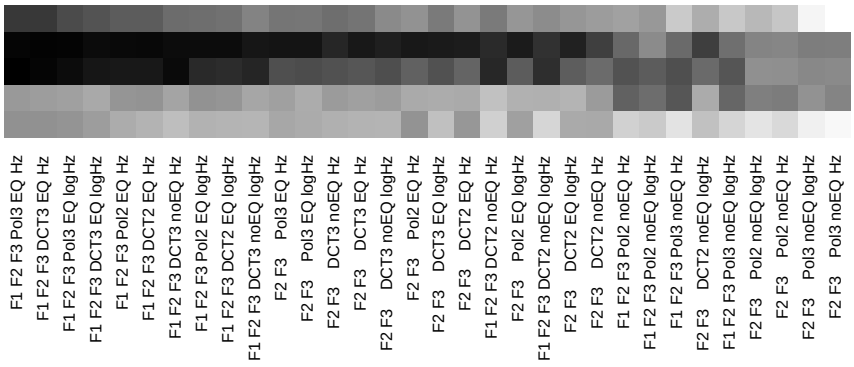


Figura 4.11: Comparación visual de rendimiento entre estrategias. Las columnas son las diferentes estrategias, y las filas son los diptongos, de arriba a abajo: /oɪ/, /aʊ/, /oʊ/, /eɪ/ y /aɪ/. Tonos más oscuros representan mejor rendimiento.

La figura 4.11 muestra una representación gráfica del rendimiento en forma de C_{llr}^{min} de las diferentes estrategias descritas en la Sección 3.3 en orden según su C_{llr}^{min} medio, calculado en base a sus resultados sobre todos los diptongos. Tonos más oscuros representan valores menores, y por lo tanto un mejor rendimiento. Cada columna representa una combinación, y están ordenadas de acuerdo a su mejor rendimiento valorado como C_{llr}^{min} en media sobre todos los diptongos. Elementos en la misma fila representan el mismo diptongo, de arriba a abajo: /oɪ/, /aʊ/, /oʊ/, /eɪ/ y /aɪ/.

Se observa que ecualizar la duración de la señal tiende a mejorar sensiblemente el rendimiento frente a duración natural y que el rango de frecuencias en escala natural generalmente ofrece un mejor rendimiento que en escala logarítmica. La ausencia del primer formante empeora en la mayoría de los casos el rendimiento frente a su presencia, y sobre las estrategias de ajuste paramétrico no se pueden extraer conclusiones claras, ya que el rendimiento de cada una dependerá de la naturaleza particular de las trayectorias en media de cada diptongo.

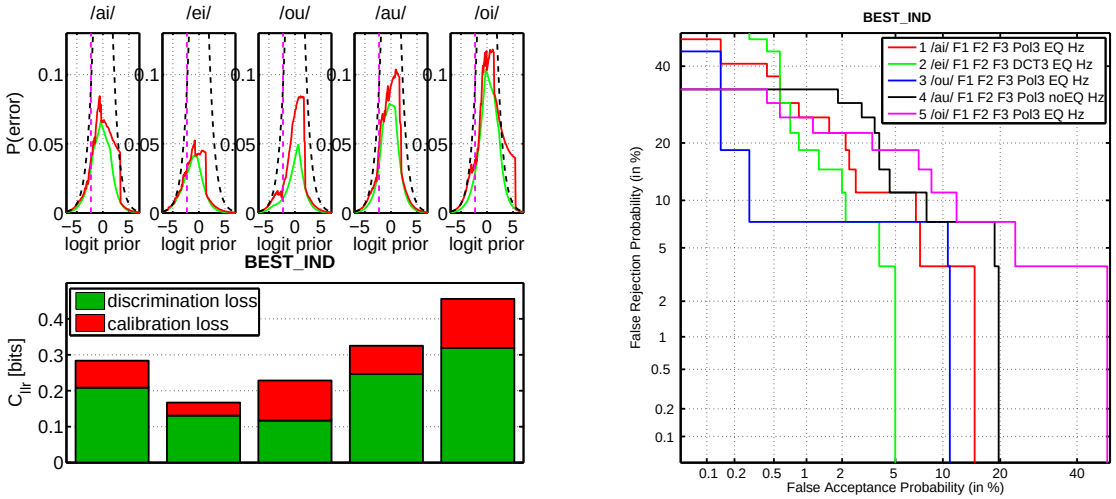
En las Figuras 4.1 a 4.9 se podía valorar el rendimiento de las diferentes estrategias sobre cada diptongo. La figura 4.11 muestra las estrategias ordenadas por rendimiento en media de cada estrategia para todos los diptongos, lo que facilita hacer observaciones más generales.

4.3.2. Rendimiento de cada estrategia para todos los diptongos

En esta sección se comparan las diferentes estrategias de extracción descritas en 3.4.2. BEST_IND seleccionaba las estrategias que ofrecían un mejor rendimiento en forma de C_{llr}^{min} para cada diptongo en particular, BEST_ALL seleccionaba la estrategia que ofrecía

un mejor rendimiento en forma de C_{llr}^{min} en media para todos los diptongos, y HUMAN_AUTO seleccionaba las estrategias que ofrecían un mejor rendimiento en forma de C_{llr}^{min} para cada diptongo usando extracción semi-automática, basándose en los resultados de [9].

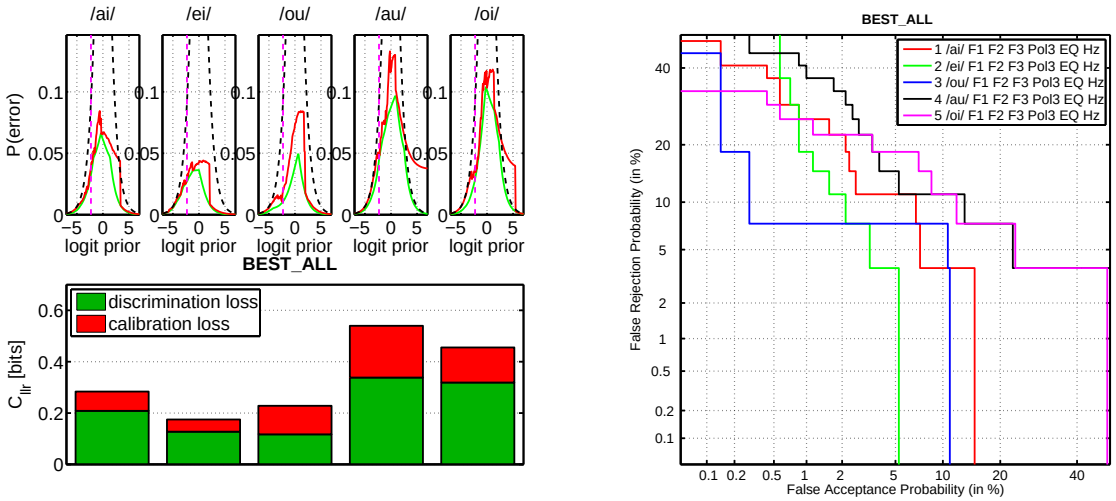
Para cada sistema, se muestra una gráfica APE y una gráfica DET. Cada una de ellas muestran de forma conjunta el rendimiento obtenido siguiendo estas estrategias de extracción para los 5 diptongos de la base de datos /aɪ/, /eɪ/, /oʊ/, /aʊ/ y /oɪ/.



(a) Curvas APE

(b) Curvas DET

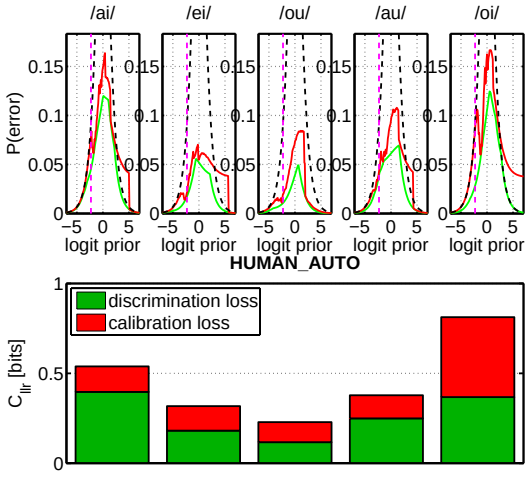
Figura 4.12: Curvas APE y DET por diptongo de la estrategia BEST_IND sobre la base de datos de *Kinoshita & Osanai 2006*.



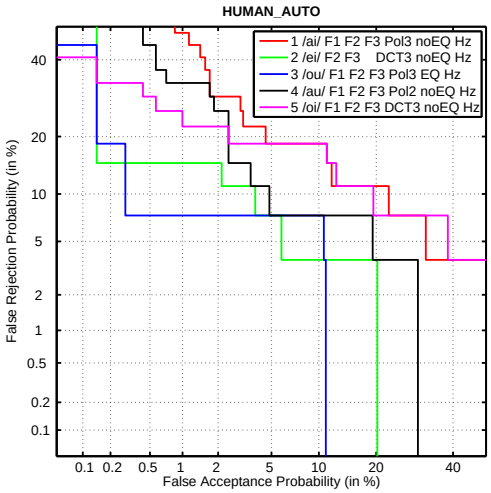
(a) Curvas APE

(b) Curvas DET

Figura 4.13: Curvas APE y DET por diptongo de la estrategia BEST_ALL sobre la base de datos de *Kinoshita & Osanai 2006*.



(a) Curvas APE



(b) Curvas DET

Figura 4.14: Curvas APE y DET por diptongo de la estrategia HUMAN_AUTO sobre la base de datos de *Kinoshita & Osanai 2006*.

Se observa que el criterio BEST_IND (Figuras 4.12(a) y 4.12(b)) ofrece los mejores resultados, como era de esperar, aunque BEST_ALL (Figuras 4.13(a) y 4.13(b)) no está demasiado por debajo en cuanto a rendimiento, máxime teniendo en cuenta que trata a todos los diptongos con el mismo criterio, lo que supone una mayor generalidad del sistema, y extrapolabilidad a nuevos diptongos en caso de aplicarla sobre otras bases de datos.

Por su parte HUMAN_AUTO tampoco se aleja demasiado en rendimiento de los dos anteriores, aunque el objetivo de esta estrategia es la comparación entre la extracción automática y semi-automática de formantes y no la obtención de los mejores resultados basados en extracción automática. Los resultados son sensiblemente inferiores a los resultados de extracción semi-automática de formantes, aunque se debe tener en cuenta el aumento del grado de automatización.

	BEST_IND		BEST_ALL		HUMAN_AUTO		HUMAN_SEMI	
Diptongo	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
/aɪ/	0.2086	0.2835	0.2086	0.2835	0.3966	0.5387	0.061	0.113
/eɪ/	0.1274	0.1670	0.1309	0.1748	0.1811	0.3174	0.063	0.095
/oʊ/	0.1170	0.2283	0.1170	0.2283	0.1170	0.2283	0.077	0.129
/aʊ/	0.2463	0.3251	0.3381	0.5397	0.2493	0.3779	0.105	0.170
/oɪ/	0.3188	0.4554	0.3188	0.4554	0.3677	0.8126	0.082	0.140

Tabla 4.3: *Comparación del rendimiento por diptongo entre todas las estrategias.*

Se observa que los resultados con extracción manual son sensiblemente mejores, alcanzando un funcionamiento muy bueno utilizando únicamente un diptongo, mientras que el mismo rendimiento de diptongos aislados presenta un rendimiento bastante inferior al tratarse de extracción automática, aunque se mantiene dentro de unos márgenes razonables.

En líneas generales se aprecia que la discriminación de los diptongos individuales es sensiblemente superior en el caso de extracción realizada o supervisada por expertos frente

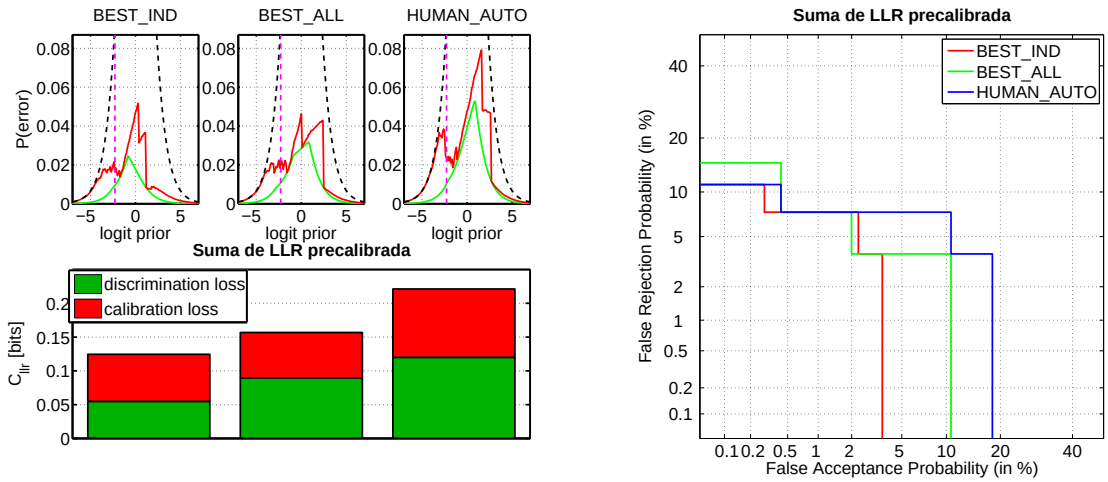
a extracción automática. Es probable que este menor rendimiento de la extracción automática esté causado por el formante F3, de difícil extracción, especialmente teniendo en cuenta que algunos sistemas semi-automáticos requieren supervisión humana a la hora de elegir la trayectoria de F3 entre varias posibilidades extraídas a través de procedimientos diferentes.

4.3.3. Fusión mediante suma precalibrada y suma postcalibrada

Tomando los resultados anteriores, en forma de LR individual para cada comparación y diptongo, tenemos para cada comparación individual cinco (5) resultados independientes, al estar cada uno basado en diptongos diferentes, de la misma comparación entre cualesquiera dos identidades de locutor.

Como se describió en la Sección 3.4.3, existen diferentes posibilidades para combinar varios LR. La más inmediata es a la suma directa de los resultados en escala logarítmica, pero es necesario aplicar un proceso de normalización o calibración en alguna fase del proceso para ofrecer resultados lo más acordes posible a todas las fuentes de información disponibles. Se puede optar entonces por aplicar el proceso de calibración a los resultados individuales y sumar resultados ya calibrados, lo que nos genera un resultado global que puede no precisar ser calibrado de nuevo, si las fuentes de información son suficientemente independientes.

Para cada criterio de selección de los descritos en la Sección 3.4.2 (BEST_IND, BEST_ALL, y HUMAN_AUTO) este proceso genera los resultados que se muestran a continuación.



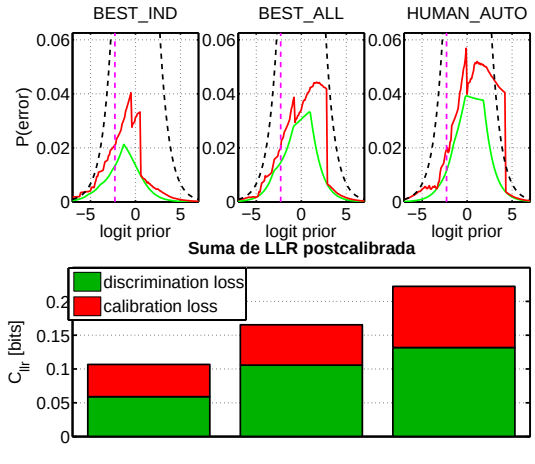
(a) Curvas APE

(b) Curvas DET

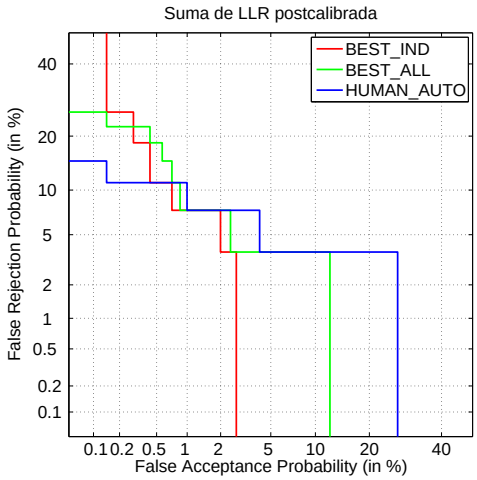
Figura 4.15: Curvas APE y DET comparativas del rendimiento ofrecido por el sistema completo aplicando para la fusión una suma precalibrada de los Log-LR de todos los diptongos (/aɪ/, /eɪ/, /oʊ/, /aʊ/ y /oɪ/) sobre la base de datos de *Kinoshita & Osanai 2006*.

Por el contrario, se puede optar por sumar en primera instancia las relaciones de verosimilitud de los que disponemos (en escala logarítmica) y aplicar a continuación un único proceso de calibración sobre el resultado global. De este procedimiento cabe esperar resultados inferiores al arrastrar el posible error de calibración inicial de cada medida, sin

embargo el tiempo de cálculo empleado para calibrar el resultado es cinco veces menor. Este proceso genera los resultados que se muestran a continuación para cada criterio:



(a) Curvas APE



(b) Curvas DET

Figura 4.16: Curvas APE y DET comparativas del rendimiento ofrecido por el sistema completo aplicando para la fusión una suma de los Log-LR de todos los diptongos postcalibrada (/aɪ/, /eɪ/, /oʊ/, /aʊ/ y /oɪ/) sobre la base de datos de *Kinoshita & Osanai 2006*.

Se observa que empleando estos procedimientos de fusión los resultados mejoran sensiblemente frente a los de los diptongos individuales, ya que fusionando los diferentes resultados individuales se da la posibilidad de que los errores causados por un mal resultado en un diptongo individual, sean solventados gracias a buenos resultados en el resto de diptongos.

Por otra parte se observa que la calibración previa a la suma en escala logarítmica ofrece un mejor resultado que la calibración a posteriori.

	PRECALIBRADO		POSTCALIBRADO	
Sistema	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
BEST_IND	0.0549	0.2696	0.0590	0.1067
BEST_ALL	0.0896	0.4165	0.1056	0.1653
HUMAN_AUTO	0.1200	0.6389	0.1317	0.2222

Tabla 4.4: *Resultados de suma logarítmica con precalibración y postcalibración.*

4.3.4. Fusión con regresión logística sobre conjuntos de 3 diptongos

Otro método de fusión propuesto en la Sección 3.4.3 es aplicar regresión logística sobre el conjunto de resultados. No obstante la fusión por regresión logística requiere un número elevado de muestras de entrenamiento cuando el número de fuentes de información a fusionar es elevado. Lo reducido de la base de datos empleada, imposibilita realizar fusión por regresión logística de la totalidad de diptongos disponibles.

Se ha procedido por tanto a efectuar combinaciones de los diptongos en grupos de a 3, que sí permiten fusión por regresión logística con la cantidad de datos disponible en esta

base de datos. Aunque los resultados no utilizan toda la información posible, si pueden reflejar la tendencia del funcionamiento de la fusión de un mayor conjunto de diptongos a la hora de trabajar sobre bases de datos mayores que sí permitan este supuesto.

A continuación se muestran los resultados de fusión con regresión logística para BEST_IND.

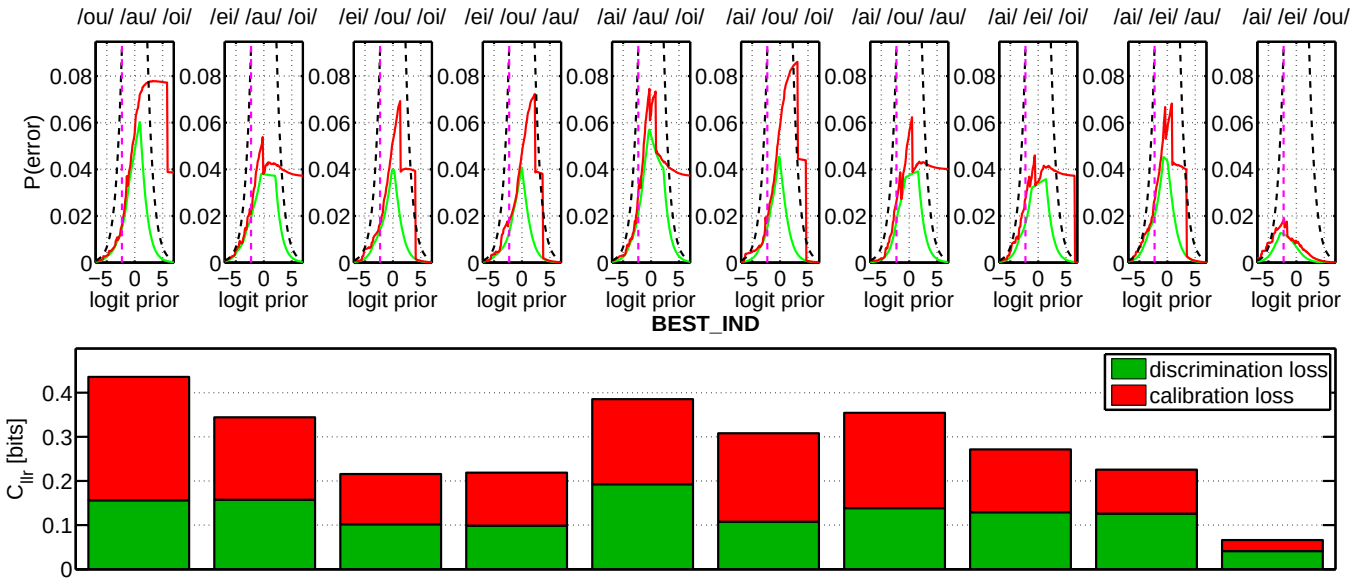


Figura 4.17: Curva APE para la estrategia BEST_IND tras la calibración y fusión en un solo paso con regresión logística de grupos de 3 diptongos.

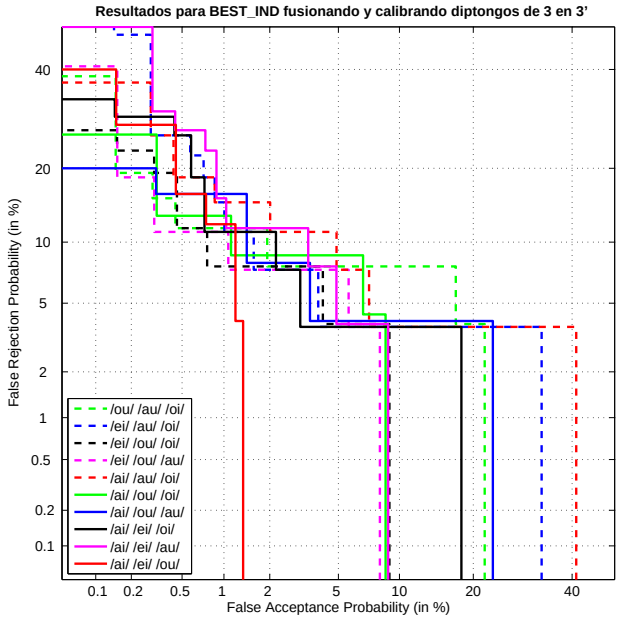


Figura 4.18: Curva DET para la estrategia BEST_IND tras la calibración y fusión en un solo paso con regresión logística de todos los grupos posibles de 3 diptongos.

4.3.5. Comparación entre estrategias de fusión

Para poder efectuar valoraciones acerca del rendimiento de los resultados obtenidos a través de fusión por regresión logística, parece razonable comparar dichos resultados con los obtenidos por otras estrategias de fusión que hagan uso de los mismos datos. Teniendo en cuenta que la extensión de la base de datos no permitió el empleo de la totalidad de la información utilizable, sino que obligó a extraer resultados selectivos en base a combinaciones de 3 diptongos de los 5 disponibles, parece razonable efectuar comparación con el resto de estrategias de fusión presentadas en la Sección 3.4.3 adaptadas para que empleen exclusivamente la misma información.

Por tanto procederemos a efectuar una comparación entre los resultados obtenidos a través de fusión por medio de regresión logística, suma precalibrada y suma postcalibrada, empleando en todos los casos combinaciones de 3 diptongos entre los 5 disponibles. Se pueden efectuar por tanto 10 comparaciones diferentes entre regresión logística y suma pre y postcalibrada. Al ofrecer todas las estrategias rendimientos similares, se ha optado por mostrar solo los resultados para BEST_IND al considerarlos suficientemente representativos.

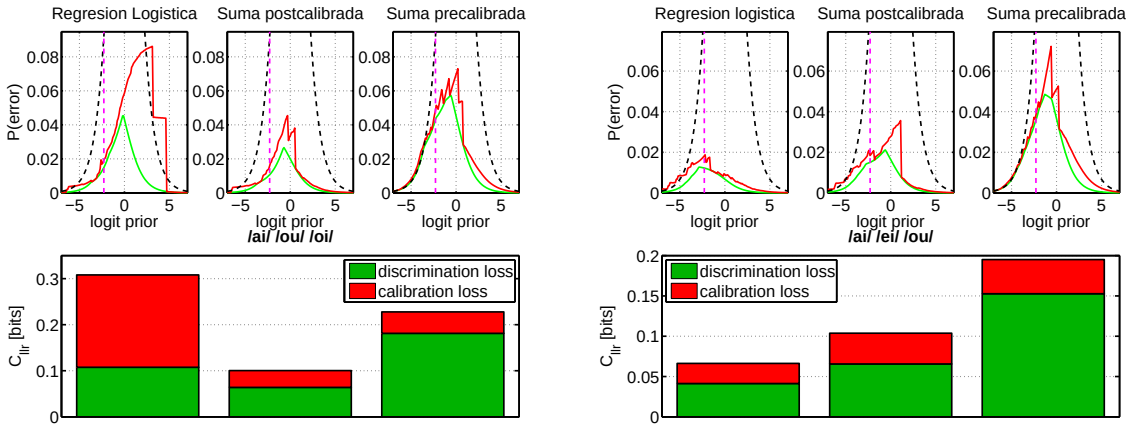
	Regresión Logística	Suma+Calibrado	Calibrado+Suma
/oʊ/ /aʊ/ /oɪ/	0.1563	**0.0858**	0.1939
/eɪ/ /aʊ/ /oɪ/	0.1575	**0.1320**	0.2053
/eɪ/ /oʊ/ /oɪ/	0.1019	**0.0727**	0.1785
/eɪ/ /oʊ/ /aʊ/	0.0989	**0.0810**	0.1658
/aɪ/ /aʊ/ /oɪ/	0.1926	**0.1377**	0.2104
/aɪ/ /oʊ/ /oɪ/	0.1076	**0.0639**	0.1812
/aɪ/ /oʊ/ /aʊ/	0.1384	**0.0938**	0.1792
/aɪ/ /eɪ/ /oɪ/	0.1291	**0.1223**	0.1900
/aɪ/ /eɪ/ /aʊ/	0.1260	**0.0942**	0.1735
/aɪ/ /eɪ/ /oʊ/	**0.0413**	0.0655	0.1527

Tabla 4.5: *Comparación de rendimiento en forma de C_{llr}^{min} obtenido por fusión de combinaciones de tres diptongos mediante regresión logística, suma postcalibrada y suma precalibrada para la estrategia BEST_IND.*

Se observa que en estas condiciones la estrategia basada en suma postcalibrada es la que ofrece en general un rendimiento superior, tal y como se puede ver en la Figura 4.19(a) de una comparación relevante a modo de muestra. Se da una excepción importante en el caso particular del conjunto de diptongos /aɪ/ /eɪ/ y /oʊ/, reflejada en la figura 4.19(b), en que el resultado obtenido empleando regresión logística es en el único caso en que es superior al obtenido por suma postcalibrada, sin embargo esto no es óbice para que dicho resultado sea mejor que los de suma postcalibrada para todas las demás combinaciones posibles de diptongos.

Esto impide la extracción de conclusiones definitivas acerca del rendimiento de cada estrategia, sin embargo puede llevar a pensar que para cantidades mayores de información, el rendimiento ofrecido por la fusión con regresión logística mejore sensiblemente frente a sumas precalibradas y postcalibradas respectivamente.

También llama la atención que, como se vio en la Sección 4.3.3, al efectuar la fusión de los cinco diptongos, el rendimiento de la fusión por suma precalibrada era ampliamente superior que por suma postcalibrada, sin embargo en los resultados de esta Sección el resultado es el contrario para todos los casos. De ello se puede deducir que la reducción de las dimensionalidad del problema eleva significativamente la relevancia del calibrado



(a) /aɪ/ /oʊ/ /oɪ/ (b) /aɪ/ /eɪ/ /oʊ/

Figura 4.19: Ejemplos de APE comparativas entre resultados obtenidos por fusión de los resultados individuales de diptongos mediante regresión logística, suma postcalibrada y suma precalibrada para la estrategia BEST_IND.

final en detrimento del calibrado previo.

4.4. Selección de características sobre base de datos de *Kinoshita & Osanai 2006*

En esta Sección se va a analizar el rendimiento del sistema propuesto de selección de características descrito en la Sección 3.5 sobre la base de datos de *Kinoshita & Osanai 2006* descrita en la Sección 4.2.1. El sistema de selección de características está basado en el tratamiento individual de características después de un ajuste paramétrico de las trayectorias formánticas. Tal y como se describe en la Sección 4.2.1 la base de datos de *Kinoshita & Osanai 2006* es una base de datos de habla microfónica inglesa por hablantes australianos.

4.4.1. Tablas de C_{llr}^{min}

Tal y como se describió en 3.5, el algoritmo comienza evaluando el rendimiento individual de cada característica (entendida como cada coeficiente individual de los que describen la curva paramétrica que mejor se ajustaba a la trayectoria real de cada formante) calculando el C_{llr}^{min} que ofrecería un sistema ficticio que sólo empleara dicha característica y cuyos resultados fueran relaciones de verosimilitud unidimensionales calculadas en base a ella.

Estos C_{llr}^{min} que representan el poder discriminativo de cada característica se ordenaban en forma de tabla para su cómodo acceso a la hora de compararlas y seleccionarlas. En las tablas 4.6, 4.7, 4.8 se reflejan estos valores empleando las estrategias BEST_IND, BEST_ALL y HUMAN_AUTO respectivamente.

La tabla 4.6 muestra la tabla de los C_{llr}^{min} calculados individualmente con la estrategia BEST_IND. Todos los diptongos están ajustados a una ecuación polinómica de grado 3 con salvedad de /aʊ/ que sigue un ajuste por DCT. De esta estrategia cabe esperar que ofrezca

	F1				F2	
Diptongo	x^3	x^2	x^1	x^0	x^3	x^2
/aɪ/	0.7337	0.7931	0.5212	0.7314	0.4701	0.5676
/eɪ/	0.7091	0.7368	0.7079	0.812	0.4633	0.6626
/oʊ/	0.6145	0.7117	0.6527	0.809	0.5476	0.5809
/aʊ/	0.7429	0.7989	0.8205	0.7967	0.5648	0.7928
/oɪ/	0.728	0.8082	0.7252	0.78	0.63	0.6641

	F2		F3			
Diptongo	x^1	x^0	x^3	x^2	x^1	x^0
/aɪ/	0.571	0.6984	0.6915	0.8607	0.8478	0.853
/eɪ/	0.5795	0.7931	0.6265	0.6186	0.6484	0.7701
/oʊ/	0.6001	0.6863	0.7279	0.7378	0.7638	0.7934
/aʊ/	0.6998	0.715	0.6518	0.6352	0.8261	0.8305
/oɪ/	0.65	0.803	0.6917	0.7452	0.6084	0.8304

Tabla 4.6: *Tabla de C_{llr}^{min} de características empleando BEST_IND*

los mejores resultados para todos los diptongos, ya que se selecciona individualmente la mejor estrategia de extracción (de forma global para los 3 formantes) para cada diptongo.

Con una primera inspección se observa que la mayoría de las características individuales de mayor poder discriminativo se concentran dentro de F2, y que el poder discriminativo de las características de F1 y F3 es en general ligeramente inferior. En concreto, la característica con mayor poder discriminativo (0.4633) es la correspondiente con el coeficiente cúbico de F2. Esta característica será la primera escogida por el sistema y sucesivamente se valorará una a una el resto de características y la influencia sobre el rendimiento del sistema que conlleva su selección.

	F1				F2	
Diptongo	x^3	x^2	x^1	x^0	x^3	x^2
/aɪ/	0.7337	0.7931	0.5212	0.7314	0.4701	0.5676
/eɪ/	0.7091	0.7368	0.7079	0.812	0.4633	0.6626
/oʊ/	0.6145	0.7117	0.6527	0.809	0.5476	0.5809
/aʊ/	0.7566	0.7976	0.8248	0.7898	0.645	0.8323
/oɪ/	0.728	0.8082	0.7252	0.78	0.63	0.6641

	F2		F3			
Diptongo	x^1	x^0	x^3	x^2	x^1	x^0
/aɪ/	0.571	0.6984	0.6915	0.8607	0.8478	0.853
/eɪ/	0.5795	0.7931	0.6265	0.6186	0.6484	0.7701
/oʊ/	0.6001	0.6863	0.7279	0.7378	0.7638	0.7934
/aʊ/	0.7245	0.7269	0.7294	0.6545	0.8215	0.8253
/oɪ/	0.65	0.803	0.6917	0.7452	0.6084	0.8304

Tabla 4.7: *Tabla de C_{llr}^{min} de características empleando BEST_ALL*

La tabla 4.7 muestra la tabla de los C_{llr}^{min} calculados individualmente empleando la estrategia BEST_ALL. Todos los diptongos están ajustados a una ecuación polinómica de grado 3, ya que BEST_ALL implica la misma estrategia global de extracción para todos los diptongos y por tanto el mismo criterio de ajuste polinómico.

Para los diptongos /aɪ/, /oʊ/ y /oɪ/ coincide exactamente con BEST_IND, por lo tanto los resultados serán los mismos. Para el resto de diptongos (/eɪ/ y /aʊ/) cabe esperar que los resultados sean ligeramente inferiores, sin embargo se gana en cuanto a automatización, ya que este sistema se podría concretar en un sistema global de extracción razonablemente eficaz para todos los diptongos, y del que cabría esperar que también funcionase de forma aceptable para diptongos nuevos, diferentes de los de esta base de datos.

	F1				F2	
Diptongo	x^3	x^2	x^1	x^0	x^3	x^2
/aɪ/	0.6825	0.8048	0.5582	0.7337	0.5731	0.5347
/eɪ/	-	-	-	-	0.7948	0.6123
/oʊ/	0.6145	0.7117	0.6527	0.809	0.5476	0.5809
/aʊ/	-	0.7224	0.806	0.6433	-	0.7762
/oɪ/	0.7138	0.7897	0.7242	0.7841	0.5617	0.6636

	F2		F3			
Diptongo	x^1	x^0	x^3	x^2	x^1	x^0
/aɪ/	0.5889	0.7192	0.6222	0.8562	0.8572	0.8208
/eɪ/	0.6634	0.576	0.7691	0.6628	0.6475	0.6959
/oʊ/	0.6001	0.6863	0.7279	0.7378	0.7638	0.7934
/aʊ/	0.7636	0.5512	-	0.8019	0.8559	0.5737
/oɪ/	0.6498	0.7549	0.6732	0.7358	0.6306	0.7832

Tabla 4.8: *Tabla de C_{llr}^{min} de características empleando HUMAN_AUTO*

La tabla 4.8 muestra la tabla de los C_{llr}^{min} calculados individualmente empleando la estrategia HUMAN_AUTO. Replicando los criterios de extracción que mejor resultado generaron en [9], los diptongos /aɪ/ y /oʊ/ están ajustados a una curva polinómica de grado 3, /aʊ/ está ajustado a una curva polinómica de grado 2, y /eɪ/ y /oɪ/ siguen una curva definida por una DCT. El objetivo de esta estrategia era la comparación con resultados semi-automáticos por parte de expertos humanos. Sin embargo al no disponer de las trayectorias extraídas no se puede comprobar el funcionamiento de selección de características con extracción semi-automática. En cualquier caso sí se puede realizar una comparación con los resultados de la Sección 4.3 obtenidos con HUMAN_AUTO y el esquema de ajuste paramétrico para evaluar la mejora introducida por la selección de características.

4.4.2. Estrategia BEST_IND

Esta Sección muestra los resultados obtenidos por la estrategia de selección de características basada en el criterio de extracciones BEST_IND. La Figura 4.20 muestra la evolución del valor de C_{llr}^{min} de medida de rendimiento del sistema basado en selección de características que emplea la estrategia de extracción BEST_IND a medida que se van añadiendo características nuevas. La curva es forzosamente monótona decreciente por definición, ya que el sistema implica la no adición de características nuevas en el supuesto de que aumente (empeore) el rendimiento del sistema medido como C_{llr}^{min}.

Se observa en la gráfica que el poder de discriminación medido como C_{llr}^{min} alcanza el cero con la adición de la 17^a característica. Un $C_{llr}^{min} = 0$ implica separación total entre comparaciones target y non-target, e implica que el sistema no es mejorable en cuanto a poder discriminativo. Aunque no es observable en la gráfica el valor de C_{llr} (que

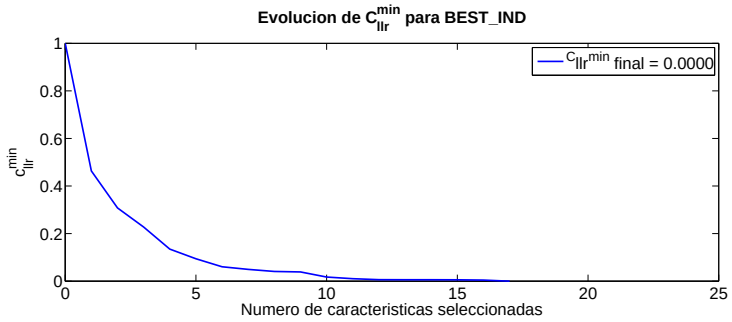


Figura 4.20: Evolución del valor de C_{llr}^{min} para la estrategia BEST_IND.

comprende el error por calibración) es de 0.0192, un valor bastante bajo representativo de un buen funcionamiento.

La Figura 4.21 muestra la tabla de selección de características. Todas las trayectorias de formantes están ajustados a una ecuación polinómica de grado 3 con salvedad de /aʊ/ que sigue una DCT. En la Figura 4.21 se reflejan las características que han sido escogidas para formar parte del sistema final, lo que permite efectuar observaciones acerca del conjunto seleccionado, analizando el comportamiento individual de cada diptongo, formante, y coeficiente.

En una primera observación, se ve rápidamente que en la gráfica predomina el blanco correspondiente a características no seleccionadas. Teniendo en cuenta que se alcanza separación total, se concluye que la selección de 17 características de 60 posibles indica que hay muchas características con poca información o para las que ésta no es aprovechable empleando el extractor automático de formantes. Esto permite un menor consumo de recursos y a la vez mantener (o mejorar) el rendimiento frente al uso de toda la información disponible. En cuanto a diptongos /oʊ/ es el que mayor número de características aporta (5), mientras que /aʊ/ es el que menos (1), reflejo del poder discriminativo de cada uno de ellos.

También se observa de manera inmediata que el formante que más información aporta es F2, y en particular el coeficiente del término cúbico x^3 de la ecuación que define la curva, mientras que F3 no aporta prácticamente ninguna característica, a pesar de típicamente F3 es altamente discriminativo. Esto es debido a la dificultad de alcanzar una alta precisión en la extracción de las trayectorias para F3 [1], siendo esta faceta la que

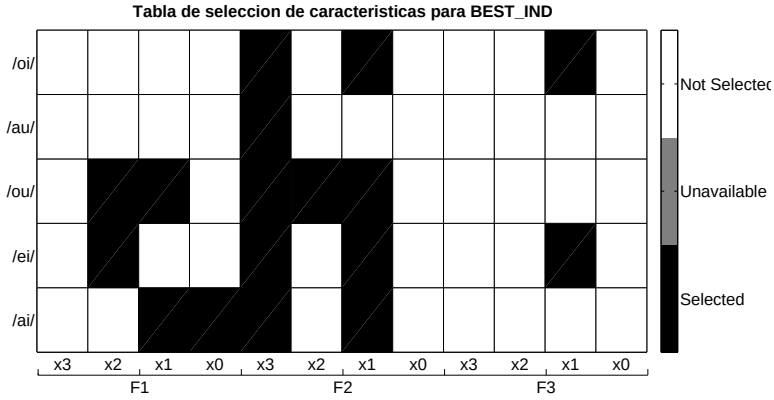


Figura 4.21: Esquema de selección de características para la estrategia BEST_IND sobre la base de datos de *Kinoshita & Osanai 2006*.

requiere mayor labor humana en extracciones semi-automáticas vigiladas por expertos.

Una gráfica DET como la de la Figura 4.22 que muestre el rendimiento del sistema en algunas etapas intermedias ofrecen una buena aproximación a la evolución del sistema con la adición de características nuevas.

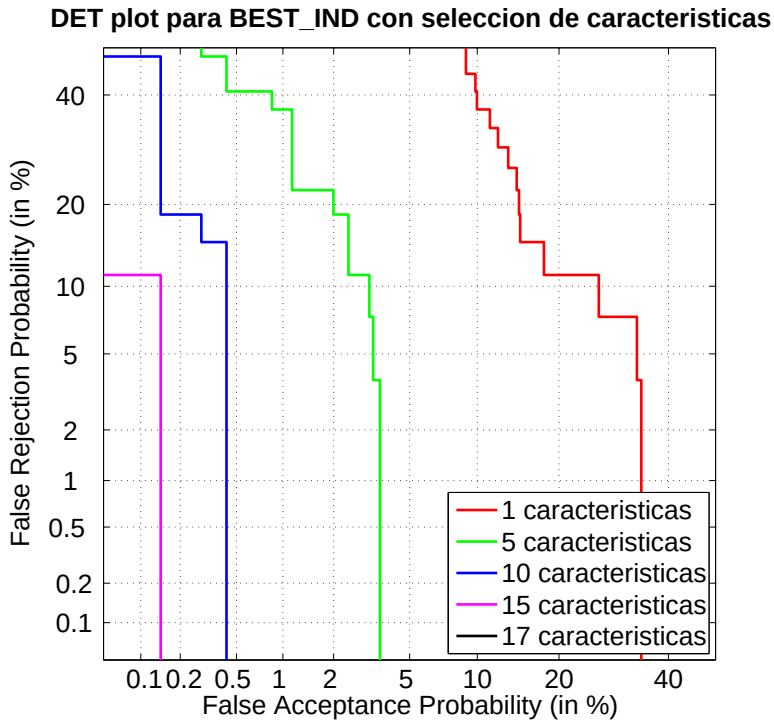


Figura 4.22: Curvas DET de la evolución del sistema de selección de características para BEST_IND sobre la base de datos de *Kinoshita & Osanai 2006*. La curva correspondiente a 17 características no es visible porque queda totalmente definida por un único puno en el origen de coordenadas (0,0).

La Figura 4.22 muestra la evolución de la curva DET que define el funcionamiento del sistema a lo largo de varias etapas, empezando con el sistema funcionando con una única característica (aunque esta es la de mejor funcionamiento entre todas las disponibles), y se muestran etapas equiespaciadas hasta la etapa definitiva del sistema, aunque en este caso no aparece ninguna curva al lograrse separación total, por lo que el sistema queda totalmente definido por un punto en el origen de coordenadas.

Es importante remarcar que con cada nueva etapa, el rendimiento del sistema no necesariamente mejora en todos los puntos de funcionamiento, aunque en la Figura 4.22 no se aprecia este fenómeno, debido al reducido número de etapas mostradas buscando una mayor claridad en la gráfica.

4.4.3. Estrategia BEST_ALL

Esta Sección muestra los resultados obtenidos por la estrategia de selección de características basada en el criterio de extracciones BEST_ALL, en el que todas las trayectorias de formantes están extraídas y ajustadas bajo las mismas condiciones buscando aumentar la generalidad. La Figura 4.23 muestra la evolución del valor de C_{llr}^{min} a medida que se van añadiendo características nuevas. Al igual que en la Sección anterior la curva es forzosamente monótona decreciente por definición.

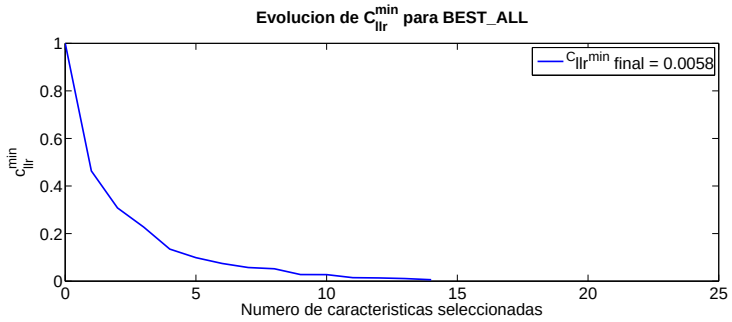


Figura 4.23: Evolución del valor de C_{llr}^{min} para la estrategia BEST_ALL.

En este caso, aplicando la estrategia BEST_ALL, no se logra separación total. Se alcanza en este caso un valor mínimo de $C_{llr}^{min} = 0{,}0058$ que no obstante sigue siendo muy bueno, especialmente teniendo en cuenta el aumento de la generalidad. Ese valor no es reducilbe con la adición de ninguna de las características disponibles al final del proceso. El valor final de C_{llr}, que también incluye el error de calibración, es $C_{llr} = 0{,}0273$.

La Figura 4.24 muestra la tabla de selección de características. En este caso, y siguiendo la estrategia BEST_ALL que busca un aumento de generalidad, la duración de todas las trayectorias ha sido previamente ecualizada, se ha mantenido la escala de frecuencia natural en Hertzios y están ajustadas a una ecuación polinómica de grado 3, tal y como se describió en la sección 3.4.2. En esta tabla se reflejan las características que han sido escogidas para formar parte del sistema final, lo que permite efectuar determinadas observaciones.

Una primera observación acerca de la Figura 4.24 permite apreciar el notable parecido con la Figura 4.21, lo que refuerza (ligeramente, al tratarse de la misma base de datos, y criterios de extracción y ajuste paramétrico no muy lejanos) las conclusiones extraídas de ella, como la importancia de F2 y la escasa aportación de F3, además de que /oʊ/ sigue siendo el diptongo que más características aporta, mientras que en este caso el diptongo /aʊ/ no sólo sigue siendo el que menos, sino que no aporta ninguna característica. El número de características elegido en este caso es de 14 de las 60 disponibles.

La Figura 4.25 que muestra la curva DET del sistema en algunas etapas intermedias para poder valorar la evolución del rendimiento del sistema con la adición de características nuevas.

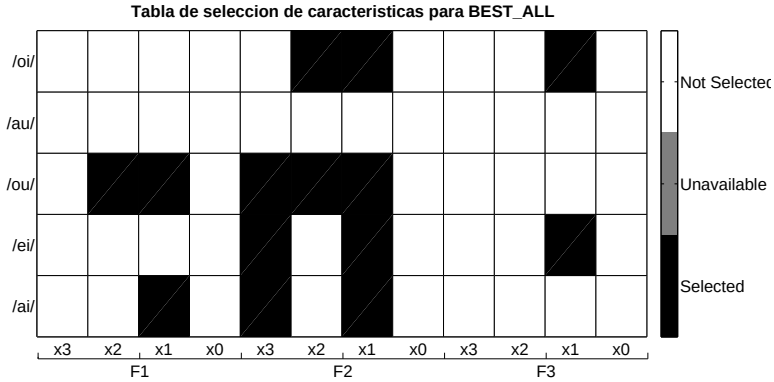


Figura 4.24: Tabla de selección de características para la estrategia BEST_ALL.

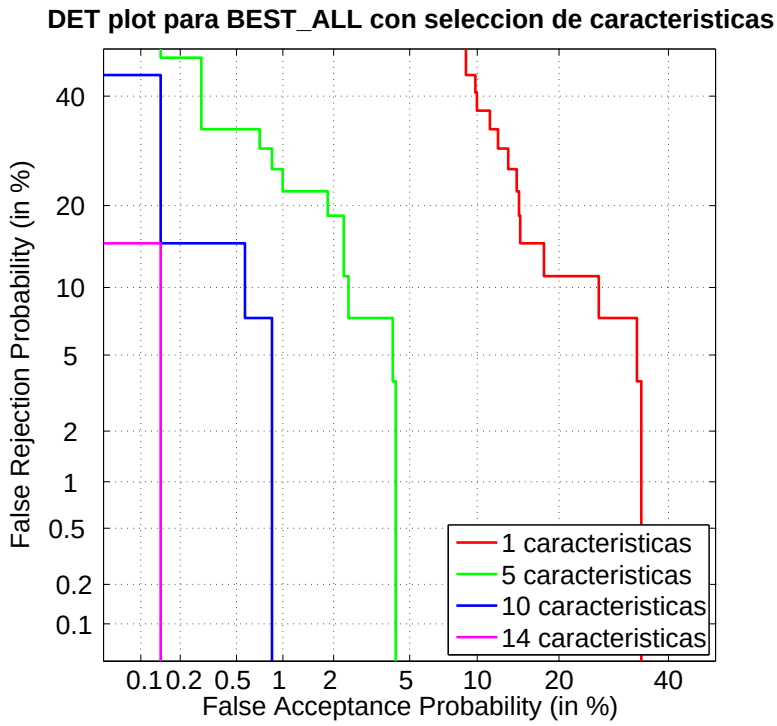


Figura 4.25: Curva DET de la evolución del sistema de selección de características para BEST_ALL.

En este caso, al no lograrse separación total, la curva correspondiente a la última etapa (14 características) define el funcionamiento del sistema final haciendo uso del conjunto de características seleccionadas. En este caso el rendimiento del sistema tampoco empeora en ningún momento para ningún punto de funcionamiento del sistema, aunque podría no ser así. De hecho si se mostraran todas las etapas sucedería con cierta frecuencia, pero la gráfica dejaría de ser suficientemente clara.

4.4.4. Estrategia HUMAN_AUTO

Esta Sección muestra los resultados obtenidos por la estrategia de selección de características basada en el criterio de extracciones HUMAN_AUTO, que replica las condiciones de extracción y ajuste paramétrico que producían mejores resultados individuales para extracción manual en [9] de cara a comparar el rendimiento de ambos. La Figura

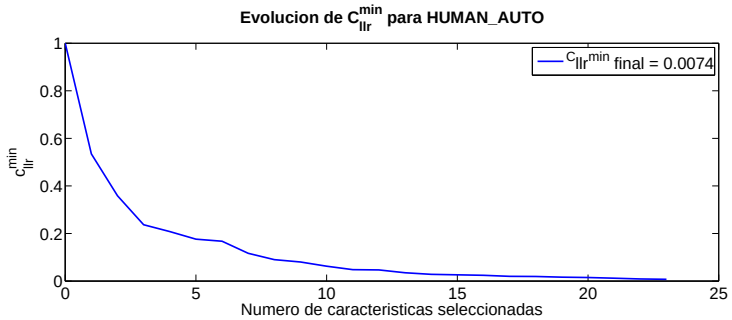


Figura 4.26: Evolución del valor de C_{llr}^{min} para la estrategia HUMAN_AUTO con selección de parámetrossobre la base de datos de *Kinoshita & Osanai 2006*.

4.26 muestra la evolución del valor de C_{llr}^{min} a medida que se van añadiendo características nuevas. La curva es monótona decreciente por definición.

En este caso, aplicando la estrategia HUMAN_AUTO, tampoco se logra separación total. Se alcanza en este caso un valor mínimo de $C_{llr}^{min} = 0{,}0074$ bastante aceptable, aunque con extracción manual y un sistema de ajuste paramétrico [9] sí se lograba separación total. El valor final de C_{llr}, comprendiendo el error de calibración, es $C_{llr} = 0{,}0225$, de nuevo muy bueno.

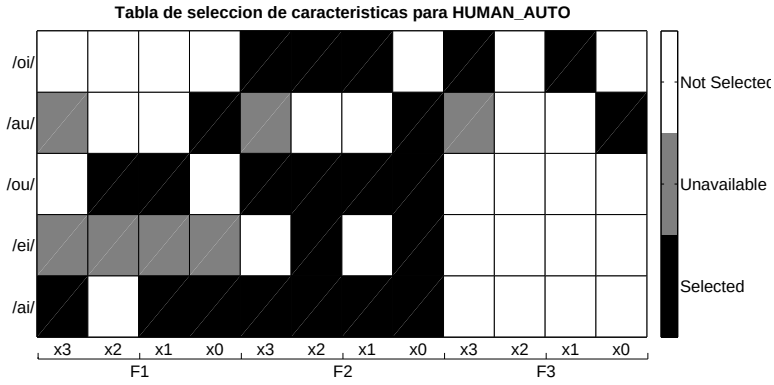


Figura 4.27: Tabla de selección de parámetros para la estrategia HUMAN_AUTO.

La Figura 4.27 muestra la tabla de selección de características. En este caso, y siguiendo la estrategia HUMAN_AUTO. Los diptongos /aɪ/ y /oʊ/ están ajustados a una curva polinómica de grado 3, /aʊ/ está ajustado a una curva polinómica de grado 2, y /eɪ/ y /oɪ/ siguen una curva definida por una DCT, tal y como se describió en la sección 3.4.2. En esta tabla se reflejan las características que han sido escogidas para formar parte del sistema final, lo que permite efectuar determinadas observaciones.

Se puede observar que la Figura 4.27 guarda algún parecido con las Figuras 4.21 y 4.24, lo que denota que características con determinada relevancia bajo los criterios BEST_IND y BEST_ALL la siguen teniendo, aunque ahora el número de características aportadas es sensiblemente mayor (23 de las 60 disponibles). Cabe destacar ahora entre los diptongos /aɪ/ /oʊ/ y /oɪ/ aportan casi toda las características y que en este caso el formante F2 adquiere todavía mayor relevancia en cuanto al número de características aportadas, mientras que F3 se queda prácticamente inédito, de nuevo a causa del inferior rendimiento de una extracción automática sin supervisión.

La Figura 4.28 muestra la curva DET del sistema en algunas etapas intermedias para poder valorar la evolución del rendimiento del sistema con la adición de características nuevas.

En este caso, al no lograrse separación total, la curva correspondiente a la última etapa, de 23 características, define el funcionamiento del sistema final haciendo uso del conjunto de características seleccionadas. Se aprecia en este caso, frente a las estrategias BEST_IND y BEST_ALL, que las etapas están más próximas entre sí, lo que significa una mejora inferior en cada paso, no obstante con un mayor número de pasos se obtienen resultados similares a los anteriores.

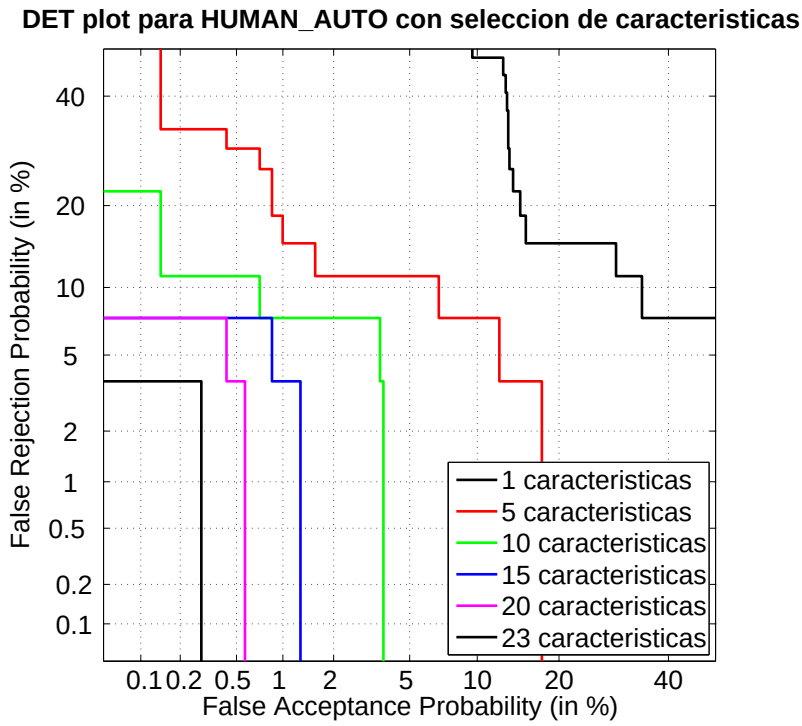


Figura 4.28: Curva DET que muestra la evolución del sistema a medida que se añaden características para la estrategia HUMAN_AUTO sobre la base de datos de *Kinoshita & Osanai 2006*.

4.4.5. Comparación entre las tres estrategias

Una APE conjunta del sistema final generado con las tres estrategias como la de la Figura 4.29 permite hacer una comparación directa del rendimiento de las mismas.

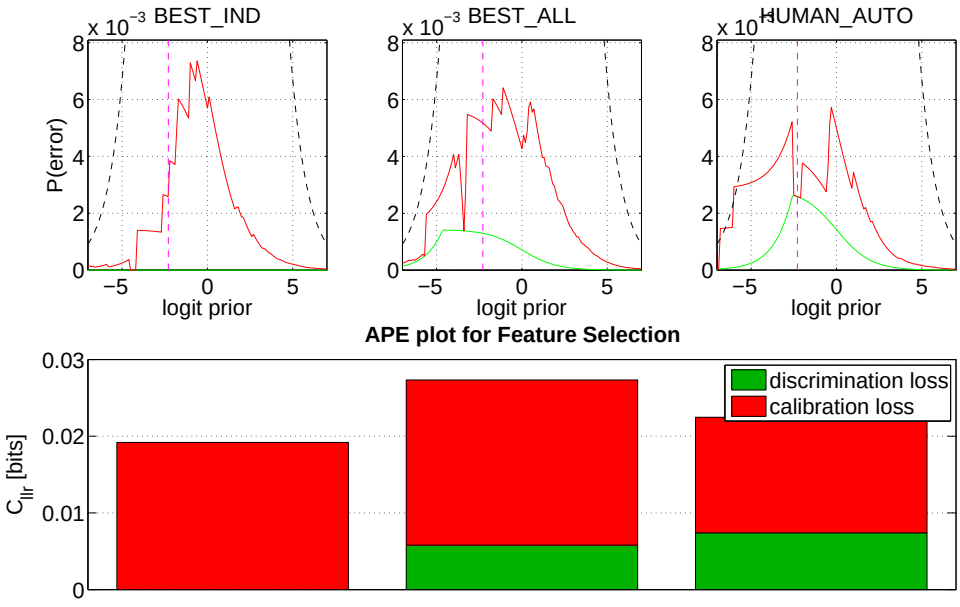


Figura 4.29: Curva APE de comparación entre las diferentes estrategias.

Como se comentó anteriormente, para BEST_IND se logra separación total $C_{llr}^{min} = 0$, aunque BEST_ALL y HUMAN_AUTO presentan valores de C_{llr}^{min} muy bajos correspondientes a rendimientos bastante buenos. Las tres estrategias también se observa que presentan valores de C_{llr} bajos, por lo que el error por calibración tampoco es muy elevado, a pesar de que en ninguna etapa de este sistema se calibra ningún resultado intermedio ni final.

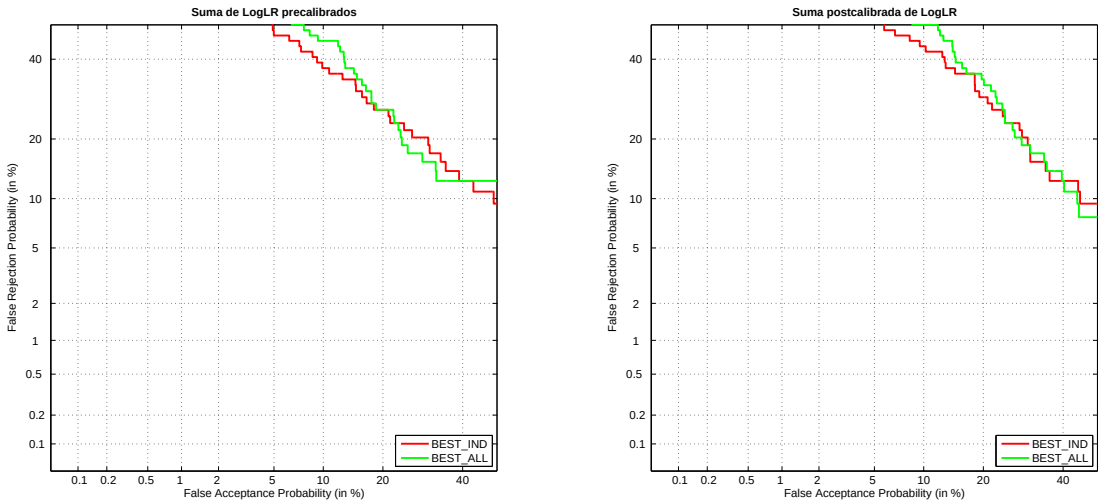
4.5. Esquema de ajuste paramétrico con seguimiento automático de formantes sobre la base de datos de *Zhang 2007*

En esta Sección se va a analizar el rendimiento del sistema de ajuste paramétrico descrito en 3.4 que hace uso del extractor automático de formantes de [10] descrito en 3.2 sobre la base de datos de *Zhang 2007* descrita en la Sección 4.2.2. Este esquema se basa en el ajuste de curvas paramétricas a las trayectorias formánticas y el uso de los parámetros que definen dichas curvas. La base de datos de *Zhang 2007* es una base de datos de hablantes de chino mandarín en conversación espontánea telefónica.

En este caso sólo se siguieron las estrategias BEST_IND y BEST_ALL, ya que HUMAN_AUTO era la réplica de otro experimento realizado con extracción semi-automática de diptongos sobre la base de datos de *Kinoshita & Osanai 2006* y en este caso pierde sentido la comparación directa.

4.5.1. Fusión mediante suma precalibrada y suma postcalibrada

Una primera vía de fusión de los resultados de los diferentes diptongos era la suma directa de los diferentes radios de verosimilitud individuales en escala logarítmica. Era necesario aplicar al proceso una etapa de calibración. Esta etapa podía ser previa a la



(a) Curva DET de la suma de LogLR precalibrados.

(b) Curva DET de la suma postcalibrada de LogLR.

Figura 4.30: Curvas DET de la fusión mediante suma precalibrada y suma postcalibrada.

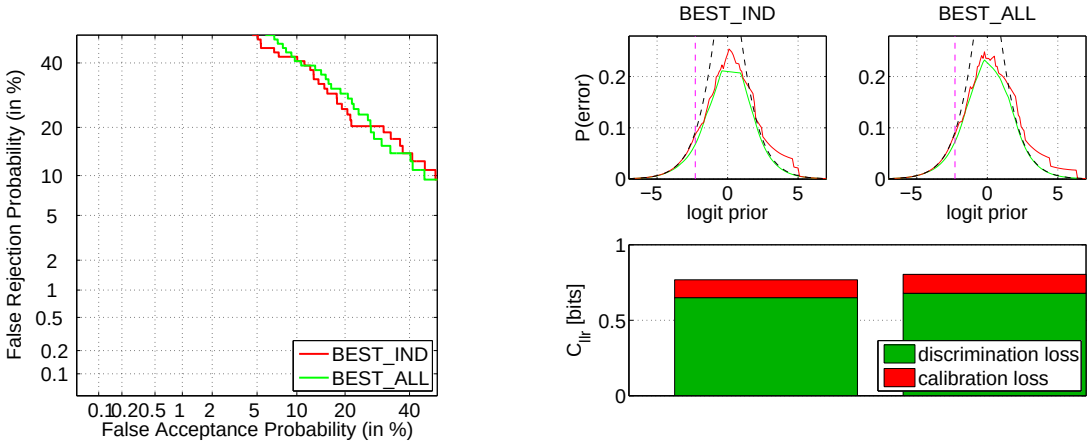
suma, es decir, de calibración individual previa de los resultados particulares, lo que evita acumular el error de calibración, o posterior a la suma, calibrando únicamente el resultado final, lo que ahorra sensiblemente tiempo de cálculo.

Los resultados ofrecen un valor de EER ligeramente superior al 20 % lo que se traduce como un rendimiento discreto. Se debe tener en cuenta al comparar con los resultados obtenidos sobre la base de datos de *Kinoshita & Osanai 2006* que las condiciones de esta eran sensiblemente más favorables que las dadas para la base de datos de *Zhang 2007* (habla controlada frente a espontánea, microfónica frente a telefónica etc.).

De nuevo se observa que la suma precalibrada funciona ligeramente mejor que la suma postcalibrada, aunque se mantienen en un margen de funcionamiento similar y el coste computacional de la suma precalibrada es bastante mayor: cada resultado precisa un número de calibraciones igual al número de diptongos, en este caso 8, frente a la suma postcalibrada que requiere tan sólo 1 calibración del resultado final, por lo que la suma precalibrada consume un 700 % más de recursos.

4.5.2. Fusión mediante regresión logística.

En este caso, la extensión de los datos disponibles sí permitió el uso de regresión logística como estrategia de fusión y calibración del sistema y por tanto se pudo fusionar toda la información disponible en un único resultado que permite efectuar una comparación directa con la suma logarítmica.



(a) Curva DET de fusión mediante regresión logística.

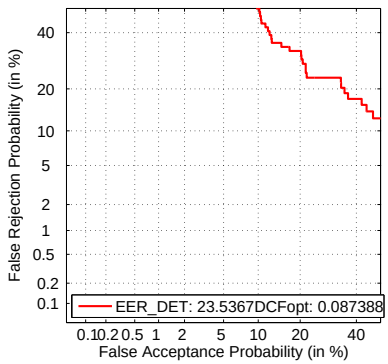
(b) Curva APE de fusión mediante regresión logística.

Figura 4.31: Curvas DET y APE de fusión mediante regresión logística.

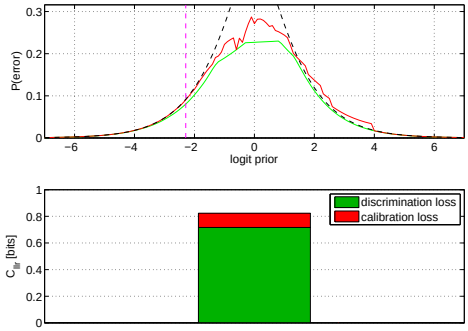
El rendimiento es ligeramente superior frente a los resultados arrojados por la suma logarítmica, aunque se mantiene en valores de EER por encima del 20 % (EER_{BEST_IND} = 21,6518 % y EER_{BEST_ALL} = 23,6359 %). El ahorro de recursos es reseñable pues en un único paso calibra y fusiona todos los resultados particulares en uno sólo, lo que combinado con el mejor funcionamiento demostrado hace que este método parezca recomendable.

4.5.3. Comparación con extracción manual basada en valores medios de los formantes.

Durante la obtención de resultados con extracción manual de formantes sobre la base de datos de *Zhang 2007* en [23] no se recurrió al ajuste paramétrico de las trayectorias de los formantes a la hora de perseguir información característica del locutor, sino que las comparaciones fueron efectuadas en base al valor medio de cada formante a lo largo del tiempo marcado como de realización de un diptongo o vocal.



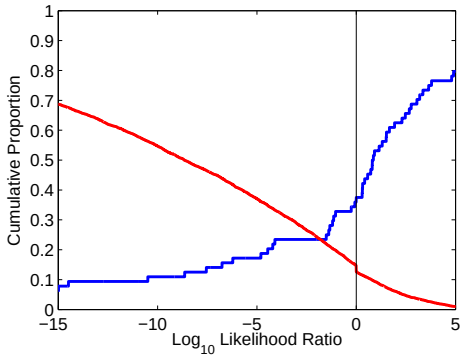
(a) Curva DET del sistema empleando valores medios de los formantes.



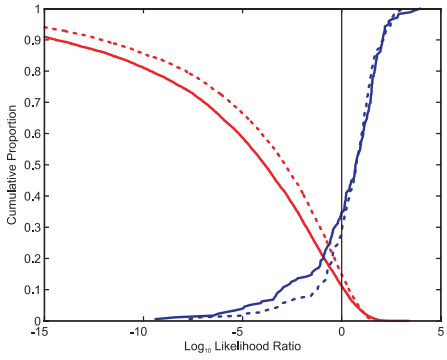
(b) Curva APE del sistema empleando valores medios de los formantes.

Figura 4.32: Curvas DET y APE del sistema empleando valores medios de los formantes extraídos automáticamente.

La Figura 4.32 evalúa los resultados de replicación completa del sistema con la inclusión de extracción automática de formantes, entendiendo que era la única comparación directa que se podía hacer entre extracción semi-automática supervisada por experto, y extracción plenamente automática de formantes. El sistema empleado para fusión de resultados es la suma logarítmica, seguido de una calibración final mediante Jackknife.



(a) Curva Tippett del sistema empleando valores medios de los formantes extraídos de manera automática.



(b) Curva Tippett de [23] empleando valores medios de los formantes extraídos de manera semi-automática. Líneas sólidas, empleando F1, F2, y F3. Líneas punteadas, usando F2 y F3 únicamente.

Figura 4.33: Curvas Tippett empleando valores medios de los formantes extraídos de manera automática y semi-automática.

La Figura 4.33 compara las curvas Tippett correspondientes a la fusión de resultados de [23] y del sistema con extracción automática de formantes, con la intención de valorar la pérdida de calidad introducida por ella, sobre la base de datos de *Zhang 2007*.

Se observa que aunque el punto de corte correspondiente al punto EER de ambas no está lejano, la extracción manual de formantes sí se traduce en un mejor rendimiento al alejarse de este punto y optar por reducir la tasa de FA o FR, frente a la otra. Algunas puntuaciones muy bajas para comparaciones target pueden suponer un problema.

4.6. Selección de características sobre la base de datos de *Zhang 2007*

En esta Sección se va a analizar el rendimiento del sistema propuesto de selección de características descrito en la Sección 3.5 basado en el tratamiento individual de características tras el ajuste paramétrico de las trayectorias formánticas, sobre la base de datos de *Zhang 2007* descrita en la Sección 4.2.2. Se trata de una base de datos de hablantes de chino mandarín en conversación espontánea telefónica.

4.6.1. Estrategia BEST_IND

La Tabla 4.9 representa los valores de C_{llr}^{min} para las diferentes características de cada diptongo y formante de los contenidos en la base de datos para la estrategia BEST_IND que selecciona para cada diptongo individualmente el tratamiento previo de señal y el ajuste paramétrico que mejor rendimiento ofrecen.

	F1				F2	
Diptongo	x^3	x^2	x^1	x^0	x^3	x^2
i1	-	0.9513	0.9479	0.9423	-	0.9344
i2	-	-	-	-	-	0.9199
i3	-	-	-	-	-	0.9493
i4	-	0.9686	0.9717	0.9611	-	0.92
y1	-	0.8837	0.9251	0.9661	-	0.8929
y2	-	-	-	-	0.9756	0.9708
y3	-	-	-	-	-	0.9319
y4	0.9778	0.9698	0.9643	0.9607	0.9481	0.9299

	F2		F3			
Diptongo	x^1	x^0	x^3	x^2	x^1	x^0
i1	0.938	0.9354	-	0.9175	0.9329	0.9092
i2	0.9138	0.9675	-	0.9242	0.9127	0.9496
i3	0.9097	0.9479	-	0.9667	0.9124	0.9432
i4	0.9373	0.9884	-	0.9484	0.954	0.9437
y1	0.9153	0.9104	-	0.9084	0.9422	0.9434
y2	0.9619	0.9641	0.9831	0.9867	0.9958	0.9406
y3	0.9702	0.9701	-	0.9254	0.958	0.9766
y4	0.9195	0.9459	0.9861	0.9848	0.9361	0.9146

Tabla 4.9: *Tabla de C_{llr}^{min} calculados individualmente con la estrategia BEST_IND para los 8 diptongos y 12 características (4 coeficientes del ajuste polinómico de 3 formantes) para utilizar en la selección de características.*

Se percibe rápidamente que los valores de C_{llr}^{min} son muy superiores a los de la base de datos de *Kinoshita & Osanai 2006*, rondando prácticamente la unidad, lo que se traduce como un rendimiento muy bajo de cada uno de ellos a nivel individual. No obstante el manejo conjunto de ellos puede mejorar el rendimiento global del sistema, aunque no sería realista esperar resultados similares a los de la base de datos de *Kinsohita & Osanai 2006*.

La Figura 4.34 muestra la evolución del valor de C_{llr}^{min} del sistema según se van añadiendo las 28 características seleccionadas en este caso.

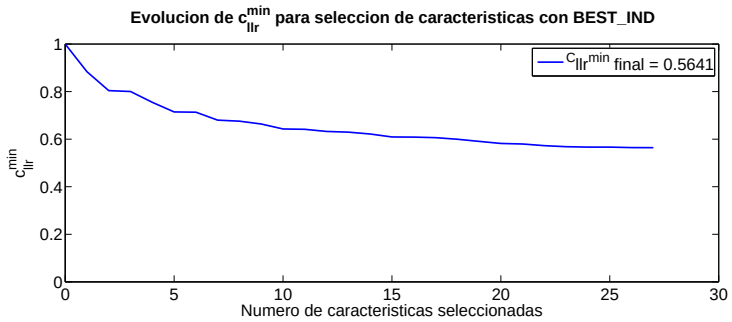


Figura 4.34: Evolución del valor de C_{llr}^{min} para la estrategia BEST_IND a medida que se van añadiendo características nuevas.

Se observa que el sistema final ofrece un rendimiento moderado, y no es capaz de mejorar el sistema más allá de un valor de $C_{llr}^{min} = 0{,}5641$. Las 28 características seleccionadas quedan definidas en la Figura 4.35.

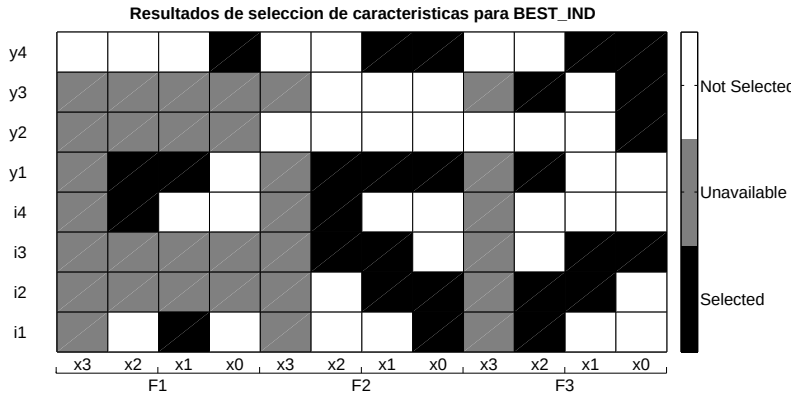


Figura 4.35: Tabla de selección de parámetros para la estrategia BEST_IND sobre la base de datos de *Zhang 2007*.

El formante F3 en este caso adquiere mayor importancia, más por el peor rendimiento de F2 en esta base de datos que por el rendimiento de F3 en sí mismo, que de hecho es pobre debido a los problemas con la extracción automática, mientras como era de esperar teniendo en cuenta que es una base de datos telefónica, F1 es significativamente menos relevante, de hecho en algunos diptongos ha sido descartado previamente por disminuir el rendimiento global del sistema y sólo han sido considerados F2 y F3.

La Figura 4.36 muestra la evolución del rendimiento del sistema a medida que se añaden características.

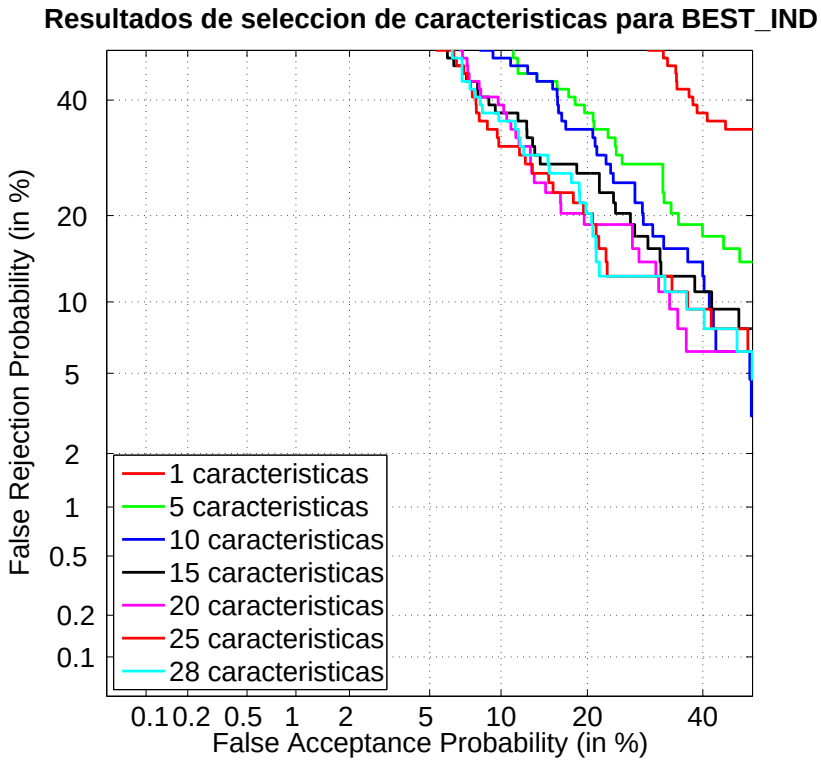


Figura 4.36: Evolución de curva DET para la estrategia BEST_IND.

La curva DET que define el funcionamiento del sistema final presenta un rendimiento peor que para la base de datos de *Kinoshita & Osanai 2006*. El espaciado entre etapas también es menor, debido al inferior poder discriminativo de las características individualmente. A pesar de mostrar etapas igualmente espaciadas que en las figuras referentes a la base de datos de *Kinoshita & Osanai 2006*, se observa que en algunos puntos de funcionamiento, una etapa ofrece peor rendimiento que la etapa anterior. Esto sucede, debido a que la mejora de calidad necesaria para aceptar una nueva característica o no, se mide en forma de C_{llr}^{min} que resume el rendimiento de toda la curva en un único número, que puede mejorar (ser inferior) a pesar de que en algún punto concreto de funcionamiento la característica nueva deteriore el sistema.

Se observa una tasa de EER del 20 % en el rango de los resultados obtenidos con extracción manual, y supera los resultados con extracción automática sin aplicar selección de características de la Sección 4.3.

4.6.2. Estrategia BEST_ALL

La Tabla 4.10 representa los valores de C_{llr}^{min} para las diferentes características de cada diptongo y formante de los contenidos en la base de datos para la estrategia BEST_ALL que selecciona el tratamiento previo de señal y el ajuste paramétrico que mejor rendimiento ofrecen de manera global en media para todos los diptongos bajo estudio.

	F1				F2	
Diptongo	x^3	x^2	x^1	x^0	x^3	x^2
i1	-	0.96654	0.97504	0.93801	-	0.93545
i2	-	0.98332	0.98156	0.96301	-	0.96233
i3	-	0.97683	0.97136	0.95144	-	0.94269
i4	-	0.9686	0.97167	0.9611	-	0.92002
y1	-	0.97105	0.94676	0.93044	-	0.90503
y2	-	0.97133	0.99263	0.98248	-	0.96354
y3	-	0.98811	0.98276	0.97533	-	0.97143
y4	-	0.94655	0.95054	0.94812	-	0.9254

	F2		F3			
Diptongo	x^1	x^0	x^3	x^2	x^1	x^0
i1	0.9407	0.94232	-	0.90463	0.927	0.92264
i2	0.96551	0.97412	-	0.94602	0.92708	0.94538
i3	0.95479	0.946	-	0.91832	0.91882	0.95186
i4	0.93733	0.98837	-	0.94837	0.95405	0.94374
y1	0.91745	0.9497	-	0.93957	0.93848	0.91554
y2	0.97411	0.96451	-	0.99555	0.99337	0.95696
y3	0.97247	0.95657	-	0.97844	0.98029	0.97607
y4	0.92594	0.93204	-	0.94225	0.9474	0.94767

Tabla 4.10: *Tabla de C_{llr}^{min} calculados individualmente con la estrategia BEST_ALL para los 8 diptongos y 12 características (4 coeficientes del ajuste polinómico de 3 formantes) para utilizar en la selección de características.*

Al igual que con BEST_IND, los valores de C_{llr}^{min} son muy superiores (por tanto peores) a los de la base de datos de *Kinoshita & Osanai 2006*, rondando prácticamente la unidad. Aunque la diferencia con respecto a BEST_IND no es muy reseñable, sí se aprecia cierto aumento en los valores de C_{llr}^{min}, de los cuales para BEST_ALL ninguno desciende de 0.9.

La Figura 4.37 muestra la evolución del valor de C_{llr}^{min} del sistema según se van añadiendo las 28 características seleccionadas en este caso.

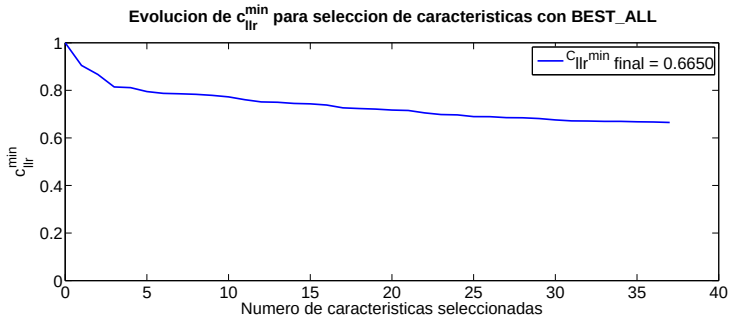


Figura 4.37: Evolución del valor de C_{llr}^{min} para la estrategia BEST_ALL sobre la base de datos de *Zhang 2007* a medida que se van añadiendo características nuevas.

El sistema final ofrece un rendimiento discreto denotado por un $C_{llr}^{min} = 0{,}6650$ que además es sensiblemente inferior al rendimiento ofrecido empleando el criterio BEST_IND, a pesar de que un número superior de características (38 frente a 28) fueron seleccionadas. Estas quedan definidas en la Figura 4.38.

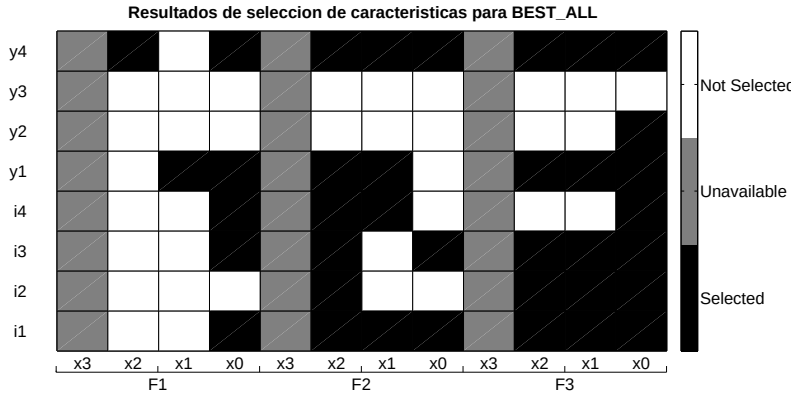


Figura 4.38: Tabla de selección de parámetros para la estrategia BEST_ALL sobre la base de datos de *Zhang 2007*.

En este caso la importancia del formante F3 es incluso superior a la de F2, que sigue siendo más relevante que F1, cuyo bajo rendimiento es debido a que el habla bajo estudio es telefónica y lo deteriora. Llama la atención que de la vocal y3 no es seleccionada ninguna característica y de y2 tan sólo una, mientras que de y4 son seleccionadas prácticamente la totalidad. De todas las /i/ se han seleccionado conjuntos de características similares, adquiriendo especial importancia F3.

La Figura 4.39 muestra la evolución del rendimiento del sistema a medida que se añaden características.

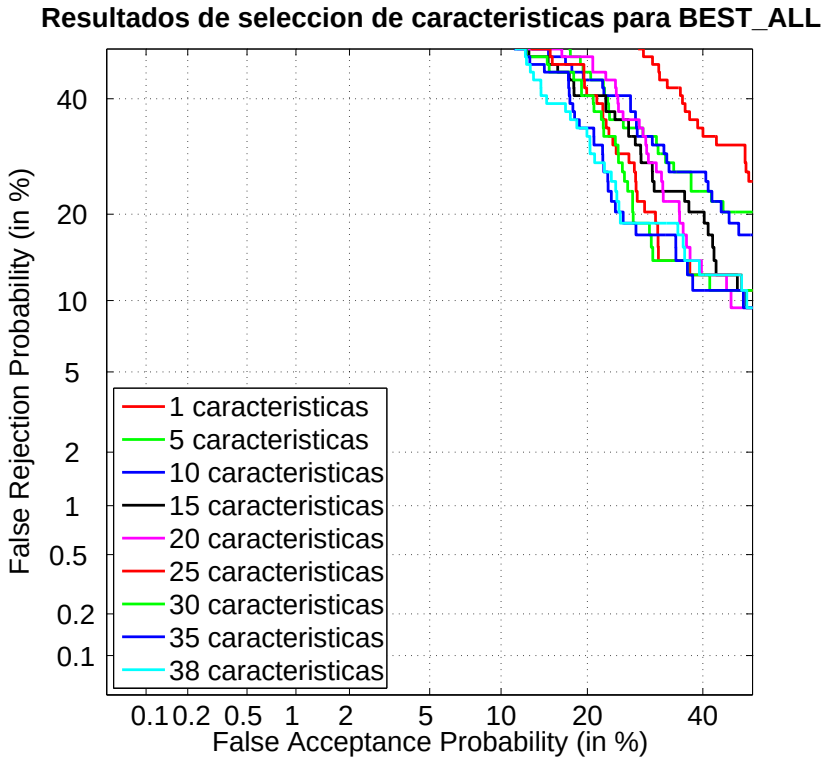


Figura 4.39: Curva DET que muestra la evolución del sistema a medida que se añaden características para la estrategia BEST_ALL sobre la base de datos de *Zhang 2007*.

En este caso nos encontramos con un rendimiento inferior al conseguido con BEST_IND. El espaciado entre etapas también es aún menor, lógicamente, al lograr mejorar en menor proporción el sistema inicial y emplear un número superior de etapas. Las líneas se entrecruzan repetidas veces por lo que en algunos puntos de funcionamiento, una la etapa ofrece peor rendimiento que la anterior. La tasa de EER es superior al 20 % por lo que en esta faceta también obtenemos peor funcionamiento que con BEST_IND.

4.6.3. Prueba de generalidad: aplicación de la estrategia BEST_ALL de la base de datos de *Kinoshita & Osanai 2006* sobre la base de datos de *Zhang 2007*

Con el objetivo de probar la generalidad de un patrón único de tratamiento previo de señal y ajuste paramétrico para trayectorias formánticas, independiente de la base de datos o condiciones de grabación, se ha efectuado una simulación del sistema completo haciendo uso de selección de parámetros sobre la base de datos de *Zhang 2007*, pero aplicando los criterios de tratamiento y ajuste obtenidos con la estrategia BEST_ALL sobre la base de datos de *Kinoshita & Osanai 2006*, de naturaleza totalmente diferente.

La Tabla 4.11 representa los valores de C_{llr}^{min} para las diferentes características de cada diptongo y formante de los contenidos en la base de datos para la estrategia BEST_AND de la base de datos de *Kinoshita & Osanai 2006*.

	F1				F2	
Diptongo	x^3	x^2	x^1	x^0	x^3	x^2
i1	0.96481	0.96712	0.97198	0.93839	0.96465	0.95589
i2	0.94119	0.95256	0.96319	0.96617	0.96483	0.94461
i3	0.98353	0.98299	0.97683	0.94371	0.97948	0.96936
i4	0.97903	0.98266	0.98529	0.96232	0.96589	0.96769
y1	0.9558	0.95549	0.94902	0.97139	0.94991	0.93447
y2	0.97892	0.98186	0.98928	0.99201	0.97409	0.9667
y3	0.98229	0.97465	0.97882	0.97638	0.93652	0.90941
y4	0.97776	0.96977	0.9643	0.9607	0.94807	0.92991

	F2		F3			
Diptongo	x^1	x^0	x^3	x^2	x^1	x^0
i1	0.93601	0.95647	0.93455	0.91073	0.86656	0.91073
i2	0.93455	0.98623	0.96437	0.96434	0.94125	0.96434
i3	0.94601	0.96266	0.95075	0.92845	0.9282	0.92845
i4	0.94899	0.95174	0.98887	0.97637	0.95861	0.97637
y1	0.92308	0.95043	0.95979	0.97559	0.97038	0.97559
y2	0.96183	0.96938	0.98407	0.98474	0.99561	0.98474
y3	0.91469	0.97218	0.98735	0.97682	0.97008	0.97682
y4	0.91954	0.94585	0.98614	0.98482	0.93609	0.98482

Tabla 4.11: *Tabla de C_{llr}^{min} calculados individualmente con la estrategia BEST_ALL de la base de datos de* Kinoshita & Osanai 2006 *para los 8 diptongos de la base de datos de* Zhang 2007 *y 12 características cada uno (4 coeficientes del ajuste cúbico de las trayectorias de 3 formantes) para utilizar en la selección de características.*

Como cabía esperar, los valores son en general ligeramente superiores a los de BEST_ALL aplicado directamente sobre esta base de datos, ya que si algún criterio su-

perara su rendimiento medio hubiera sido seleccionado como BEST_ALL en su lugar, aunque la diferencia es prácticamente nula. La Figura 4.40 muestra la evolución del valor de C_{llr}^{min} del sistema según se van añadiendo las 43 características seleccionadas en este caso.

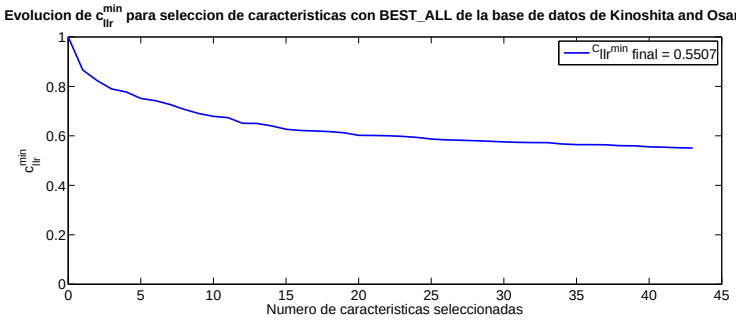


Figura 4.40: Evolución del valor de C_{llr}^{min} para la estrategia BEST_ALL de la base de datos de *Kinoshita & Osanai 2006* a medida que se van añadiendo características nuevas.

El sistema alcanza un valor de $C_{llr}^{min} = 0{,}5507$, inferior (mejor rendimiento) incluso al alcanzado con la estrategia BEST_IND, supuestamente ideada para obtener los mejores resultados. Parece por tanto posible caer en mínimos relativos en forma de valles de C_{llr}^{min} que no representen el mejor rendimiento posible del sistema (mínimo absoluto). Es posible que para un determinado ajuste, aunque no sea el mejor, alguno de sus componentes sí funcione de manera especialmente buena y su selección resulte más ventajosa, a pesar de que el rendimiento en media del ajuste baje debido a características poco discriminativas, que posteriormente serán descartadas y no pasarán a formar parte del sistema.

Al margen, unas ciertas condiciones de generalidad de los diptongos pueden ser beneficiosas para el sistema aunque contravengan los criterios previos de selección de ajuste. Por tanto un ajuste genérico (BEST_ALL de la base de datos de *Kinoshita & Osanai 2006*) puede convertirse, como en este caso, en la estrategia que genere mejores resultados.

Empleando este ajuste, han sido seleccionadas las 43 características que se muestran en la Figura 4.38.

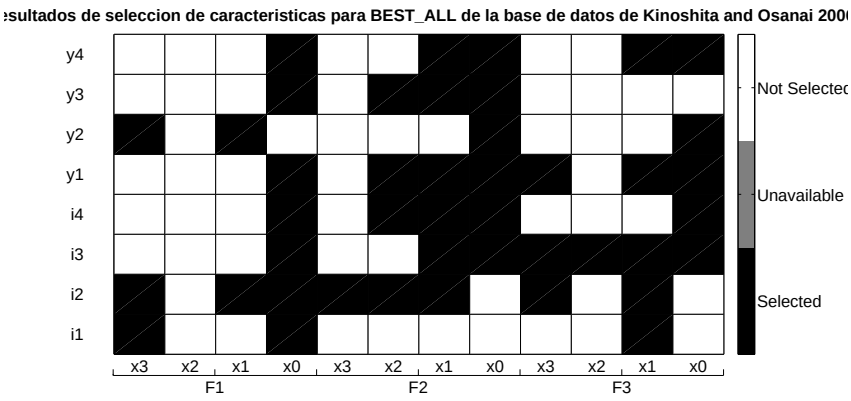


Figura 4.41: Tabla de selección de parámetros sobre la base de datos de *Zhang 2007* para la estrategia de selección BEST_ALL de la base de datos de *Kinoshita & Osanai 2006*.

F2 vuelve a ser el formante que más características aporta, mientras que F3 pierde mucha relevancia con respecto a la que tenía con los criterios BEST_IND y BEST_ALL

de la propia base de datos de *Zhang 2007*, casi equiparándose a F1, del que sin embargo es importante reseñar la importancia del coeficiente correspondiente al término independiente.

La Figura 4.42 muestra la evolución de la curva DET del sistema a medida que se añaden características.

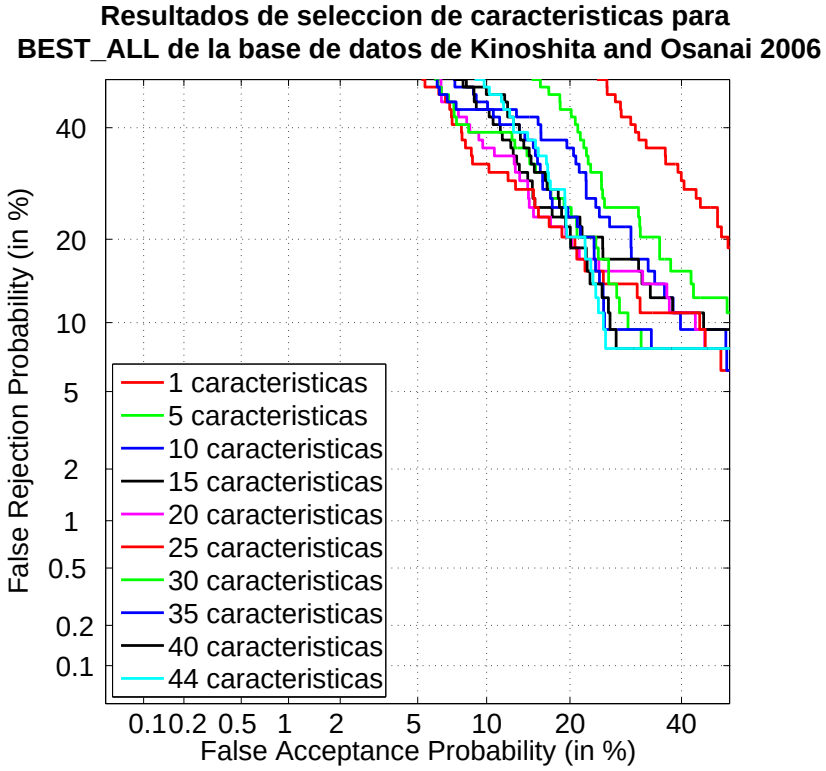


Figura 4.42: Curva DET que muestra la evolución del sistema a medida que se añaden características para la estrategia BEST_ALL de la base de datos de *Kinoshita & Osanai 2006*.

La tasa de EER ronda el 20 % logrando también en este caso un rendimiento superior al ajuste paramétrico de la Sección 4.3, y similar al obtenido aplicando extracción formántica manual. Además que los resultados sean, no sólo comparables sino superiores, a los logrados con BEST_IND, indican que ese criterio parece funcionar de forma aceptable para un amplio rango de diptongos o vocales diferentes (fue seleccionado en base a diptongos diferentes obtenidos en condiciones diferentes).

5

Conclusiones y trabajo futuro

5.1. Introducción al Capítulo

A lo largo de este proyecto se han presentado dos sistemas completos de reconocimiento forense de locutor, basados respectivamente en un esquema de ajuste paramétrico con extracción automática de formantes, y un sistema basado en selección de características. Además cada uno de ellos ha sido evaluado sobre dos bases de datos de naturalezas muy diferentes (*Kinoshita & Osanai 2006* de habla microfónica controlada y *Zhang 2007* de habla espontánea con codificación GSM).

En este Capítulo, se procede a analizar de forma global los resultados de cara a extraer conclusiones valiosas del desarrollo y funcionamiento de cada uno de estos sistemas.

Adicionalmente se hablará del posible trabajo futuro, detallando posibles mejoras al trabajo realizado, ampliaciones del mismo, o comentando la posibilidad de continuar evaluando su rendimiento en nuevas bases de datos, más amplias o más realistas.

5.2. Conclusiones

En este proyecto se han presentado dos sistemas completos, uno de ellos replica el esquema de ajuste paramétrico seguido en [9], reemplazando la extracción semi-automática de formantes, por extracción plenamente automática empleando la herramienta descrita en [10] sobre bases de datos previamente etiquetadas por expertos humanos.

El segundo sistema amplia el anterior con la aplicación de esquema nuevo que, tras el ajuste polinómico, sigue un planteamiento basado en la individualización de características en lugar de emplearlas de manera conjunta para cada formante. En esta sección se sacarán conclusiones en base a los resultados obtenidos.

5.2.1. Esquema de ajuste paramétrico utilizando seguimiento automático de formantes

A lo largo de este proyecto se ha empleado el extractor automático de formantes de [10] y se han efectuado diversas pruebas de su funcionamiento, realizando cuando ha sido posible comparaciones directas con sistemas similares que hayan utilizado extracción semi-automática de formantes o supervisada por humanos.

Los resultados del sistema completo de ajuste paramétrico empleado el extractor automático de formantes sobre la base de datos de *Kinoshita & Osanai 2006* han ofrecido valores de rendimiento buenos, a pesar de no llegar a alcanzar la separación total entre comparaciones target y non-target, hito sí alcanzado con extracción semi-automática en [9], aunque el rendimiento de los sistemas no está muy distante.

El nivel de resultados disminuye al aplicar el mismo método y protocolo sobre la base de datos de *Zhang 2007*, de tintes más realistas, es decir, de cualidades más cercanas a lo que cabría esperar en una situación real, o incluso más duras, lo que contrasta con las condiciones casi ideales de la base de datos de *Kinoshita & Osanai 2006*.

Los resultados para la base de datos de *Zhang 2007* han ofrecido un rendimiento discriminativo sensiblemente inferior (debido a la naturaleza de las grabaciones: habla espontánea GSM) aunque siguen estando en márgenes aceptables, comparándolo con trabajos similares basados en extracción semi-automática sobre el mismo cuerpo de muestras como [23].

Se puede concluir que el extractor automático de formantes de [10] funciona de manera generalmente aceptable, aunque como cabía esperar es inferior a la extracción manual o semi-automática (supervisada) de formantes. Sin embargo, el aumento en el grado de

automatización, y la muy moderada degradación relativa de los resultados, hacen que la opción de extracción de formantes sin necesidad de labor o supervisión humana cobre importancia a la hora de enfrentarse a conjuntos de grabaciones extensas, pues se agilizaría enormemente el tiempo necesario empleado y no sería necesaria la labor constante de un experto pues gran parte del trabajo sería automático.

La comparación entre los resultados automáticos y semi-automáticos para las bases de datos de *Kinoshita & Osanai 2006* y *Zhang 2007* respectivamente, permiten observar que la inferior calidad de las grabaciones no afecta en exceso al rendimiento del extractor, ya que la diferencia relativa de rendimiento frente a resultados con extracción semi-automática no sufre variaciones significativas.

5.2.2. Esquema de selección de características

Como novedad se ha planteado un esquema de selección de características, que en una etapa posterior al ajuste paramétrico, diferencia entre coeficientes tomándolos como características propias (frente al tratamiento conjunto de todos ellos como datos multivariados) y selecciona para su integración en el sistema sólo aquellos que lo mejoran, empezando por los de mejor rendimiento. Este esquema realiza, además, la fusión de los coeficientes seleccionados.

Aplicando este esquema sobre la base de datos de *Kinoshita & Osanai 2006* se obtuvieron unos resultados reseñables, logrando en el mejor de los casos, correspondiente con la estrategia denominada BEST_IND, separación absoluta ($C_{llr}^{min} = 0$) y en el resto de casos valores bastante buenos (C_{llr}^{min} por debajo de la centésima de unidad), y en todo caso errores de calibración muy bajos (cerca de 2 centésimas de unidad) pese a que no se calibra en ninguna etapa del sistema. En el caso de BEST_IND ($C_{llr} = 0{,}0192$) incluso se superaron los resultados logrados en [9] ($C_{llr} = 0{,}056$), en el que el experto intervenía en la extracción de formantes.

De nuevo sobre la base de datos de *Zhang 2007* se obtuvieron resultados peores, rondando siempre en el mejor de los casos EER del 20 %, lo que en cualquier caso mejora el rendimiento del esquema de ajuste paramétrico con el mismo extractor automático de formantes y que se sitúa en valores cercanos a la literatura disponible [23] que hace uso de esta base de datos y extracción semi-automática.

5.3. Trabajo futuro

Sobre este proyecto es factible realizar una serie de ampliaciones, especialmente interesantes aquellas destinadas a seguir aumentando el grado de automatización de los bloques del sistema, siempre con el objetivo en mente de poder aplicarlo (con la menor necesidad de labor humana adicional posible) sobre bases de datos de grandes dimensiones.

Una total automatización haría irrelevante la dimensión del material bajo estudio, al no requerir labor ni supervisión por parte de un experto humano. Esto permitiría más datos en menos tiempo y por tanto se traduciría en un funcionamiento más robusto estadísticamente a la hora de emitir un resultado.

Por otra parte, también puede ser productivo seguir efectuando pruebas sobre nuevas bases de datos diferentes, de cara a valorar el funcionamiento de los métodos descritos a lo largo de este volumen sobre materiales de diferentes condiciones.

En este aspecto, se ha de considerar que en este proyecto se han empleado bases de datos previamente segmentadas y etiquetadas de forma manual, lo que se convierte en un requisito estricto a la hora de efectuar más pruebas. No obstante, el ya mencionado

aumento del grado de automatización puede estar enfocado a eludir esta barrera automatizando la segmentación, lo que sí permitiría finalmente un estudio sobre cualquier base de datos genérica evitando el requisito previo de un alto número de horas de laboratorio enfocadas al acotado y etiquetado de material relevante.

5.3.1. Aumento del grado de automatización

A la hora de efectuar ampliaciones a los métodos, parece inmediato intentar sustituir la única fase del sistema que requiere supervisión humana. En esta fase se procedía de forma manual a localizar y segmentar los diptongos contenidos en el material de la base de datos, y a etiquetar de forma manual la vocal o diptongo pronunciado.

Por ejemplo una mejora sustancial consistiría en el embebido en el sistema de un proceso de reconocimiento de texto que pudiera facilitar la localización y extracción de forma automática de material de interés sobre audios de mayor longitud, a través de una definición previa de las realizaciones que se desee comparar, pudiendo llegarse al extremo de utilizar todos los diptongos, vocales o sílabas disponibles en la locución.

De manera ideal, este proceso devolvería para toda región de interés, las marcas temporales, o directamente la forma de amplitud de onda, o incluso los valores numéricos resultado del método de ajuste que se vaya a emplear, si este está definido previamente. También devolvería un texto o identificador que defina inequívocamente el contenido de la región para compararlo únicamente con material del mismo tipo o definido como compatible.

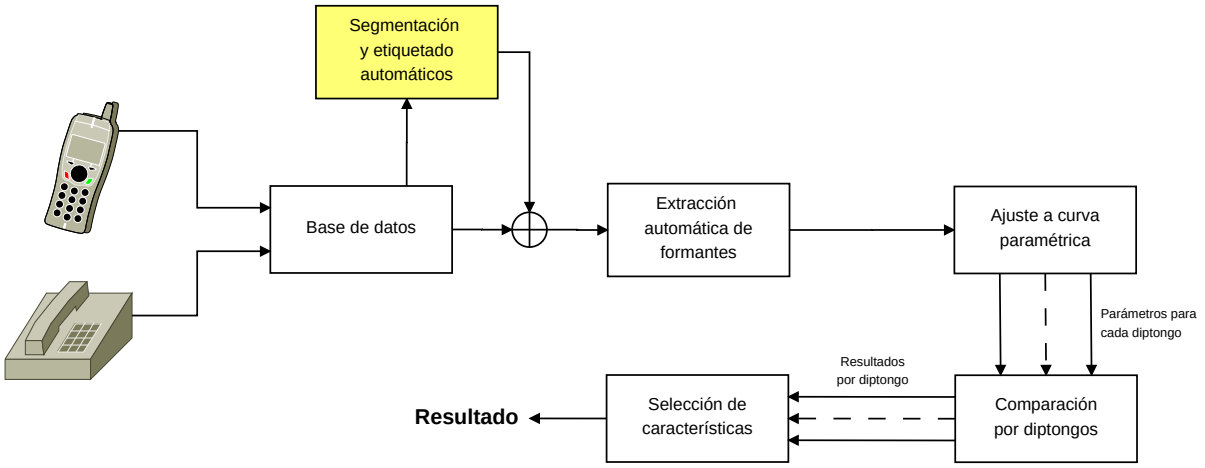


Figura 5.1: Esquema de un sistema forense de reconocimiento de locutor basado en ajuste paramétrico con segmentación, extracción y etiquetado de formantes automáticos

La Figura 5.1 muestra un posible esquema de un sistema completamente automático al suprimir totalmente la necesidad de intervención humana, integrando un módulo que segmente y etiquete de manera automática el material de interés sobre la base de datos.

5.3.2. Pruebas sobre otras bases de datos

En este proyecto, se han efectuado pruebas sobre las bases de datos de *Kinoshita & Osanai 2006* y *Zhang 2007*. Con ello se ha pretendido valorar el funcionamiento en diferentes condiciones, tanto de idioma (inglés australiano frente a chino mandarín), como de naturaleza de las muestras (habla recitada frente a espontánea, microfónica y sin compresión frente a GSM etc.).

No obstante, los individuos que forman parte de ambas son de un perfil parecido, ya que en ambos casos las muestras fueron obtenidas sobre varones cuyas edades oscilan entre 19 y 63 años para *Kinoshita & Osanai 2006* y entre 19 y 23 años para *Zhang 2007*. Esto deja fuera de la simulación diferentes perfiles, como varones de edad avanzada, niños (perfil quizás menos relevante al tratarse de reconocimiento forense de locutor), y especialmente mujeres de cualquier edad.

Sería por tanto de especial interés una prueba de los métodos descritos sobre bases de datos femeninas. Además si el idioma de la base de datos y las condiciones de obtención de muestras coinciden con los de alguna de las bases de datos simuladas en este volumen, sería posible efectuar una comparación directa de cómo el género de los individuos bajo estudio afecta al rendimiento del sistema.

También sería interesante en particular una evaluación sobre bases de datos Nist [50] de altísima variabilidad.

5.3.3. Diversificación de la información extraída

Al tratarse de un sistema automático, se podría ampliar la cantidad de información analizada en el material de estudio, añadiendo algunas características discriminativas por locutor como la frecuencia fundamental, variación de energía etc. de cualquier clasificación de las descritas en la Sección 2.2 (auditivas/acústicas y lingüísticas/no lingüísticas). Esto permitiría añadir al sistema el análisis de más información que pueda resultar relevante a la hora de identificar locutores en material dubitado. Se podría evaluar el rendimiento individual o conjunto con el sistema de características de otro nivel, o si se ha implementado un reconocedor de texto, proceder al análisis de usos infrecuentes o incorrectos del lenguaje a alto nivel, o cualquier otra cualidad que pueda entenderse como útil (o susceptible de ser útil) en reconocimiento forense de locutor.

5.3.4. Fusión con resultados obtenidos por otros métodos

Si se logra disponer de resultados obtenidos por procedimientos de naturaleza diferente o compatible (por ejemplo resultados de sistemas automáticos basados en coeficientes cepstrales) y se cumplen ciertos criterios de independencia entre los grupos de resultados, se puede proceder a la fusión para obtener un único resultado final para cada comparación que englobe toda la información analizada, contenida en cada resultado.

Esta fusión se puede afrontar integrando una implementación del método que genere estos resultados dentro del propio sistema, de tal manera que se puedan efectuar evaluaciones completas sobre bases de datos sin depender de resultados de otras simulaciones independientes.

Bibliografía

[1] A. de Castro, D. Ramos, and J. Gonzalez-Rodriguez, "Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking," in *Proc. of Interspeech*, 2009.

[2] P. Rose, *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, 2002.

[3] F. Nolan, "Speaker recognition and forensic phonetics," in *The Handbook of Phonetic Sciences*, WJ. Hardcastle and J. Laver, Eds., pp. 744–767. Oxford: Blackwell, 1997.

[4] H. J. Künzell, "Current approaches to forensic speaker recognition," in *Proc. ESCA Workshop on Automatic Speaker Recognition*, Martigny (Switzerland), 1994, pp. 135–141.

[5] D. Ramos, *Forensic evaluation of the evidence using automatic speaker recognition systems*, Ph.D. thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, 2007, Available at http://atvs.ii.uam.es.

[6] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, pp. 460–475, 1976.

[7] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Forensic identification reporting using automatic speaker recognition systems," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, 2003, vol. 2, pp. 93–96.

[8] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 159–191, 2006.

[9] G. S. Morrison, "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories," *Journal of the Acoustical Society of America*, vol. 125, no. 4, April 2009, In press.

[10] D. Rudoy, D. N. Spendley, and P. J. Wolfe, "Conditionally linear Gaussian models for estimating vocal tract resonances," in *Proc. of Interspeech*, Antwerp, Belgium, 2007, pp. 526–529.

[11] N. G. Ushakov, "Density of a probability distribution," in *Hazewinkel, Michiel, Encyclopaedia of Mathematics,*. 2001, Kluwer Academic Publishers.

[12] P. Rose, "The technical comparison of forensic voice samples," in *Expert Evidence, Issue 99*, I. Freckelton and H. Selby, Eds., pp. 3061–3062. Thomson Lawbook Company, Sydney, 2003.

[13] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[14] M. Khodai-Joopari, *Forensic Speaker Analysis and Identification by Computer: A Bayesian approach anchored in the cepstral domain*, Ph.D. thesis, University of New South Wales, Australia, 2006.

[15] D. A. Reynolds, J. P. Campbell, W. M. Campbell, R. B. Dunn, T. P. Gleason, D. A. Jones, T. F. Quatieri, C. B. Quillen, D. E. Sturim, and P. A. Torres-Carrasquillo, "Beyond cepstra: Exploiting high-level information in speaker recognition," in *Proc. Workshop on Multimodal User Authentication*, Santa Barbara, California, USA, 2003, pp. 223–229.

[16] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics*, vol. 53, pp. 109–122, 2004, with corrigendum 665-666.

[17] T. M. Nearey, P. F. Assmann, and J. M. Hillenbrand, "Evaluation of a strategy for automatic formant tracking," *J. Acoustical Soc. Amer.*, 2002.

[18] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Berlin, Springer Verlag, 1976.

[19] G. S. Morrison, "Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of australian english /ai/," *International Journal of Speech Language and the Law*, vol. 15, no. 2, pp. 249–266, 2008.

[20] Y. Kinoshita and T. Osanai, "Within speaker variation in diphthongal dynamics: What can we compare?," *Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand*, 2006.

[21] T. Kamada, N. Minematsu, T. Osanai, H. Makinae, and M. Tanimoto, "Speaker Verification in Realistic Noisy Environment in Forensic Science," *IEICE Trans Inf Syst*, vol. E91-D, no. 3, pp. 558–566, 2008.

[22] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-case database in spanish," in *Proc. of Interspeech*, Brisbane, Australia, September 2008, pp. 1493–1496.

[23] C. Zhang, G. S. Morrison, and P. Rose, "Forensic speaker recognition in chinese: A multivariate likelihood ratio discrimination on /i/ and /y/," in *Proc. of Interspeech*, 2008, pp. 1937–1940.

[24] P. Rose, "The intrinsic forensic discriminatory power of diphthongs," in *Proc. of 11th Australian International Conference on Speech Science and Technology*, 2006, (submitted).

[25] P. Rose, Y. Kinoshita, and T. Alderman, "Realistic extrinsic forensic speaker discrimination with the diphthong /ai/," *Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand*, 2006.

[26] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech and Language*, vol. 22, no. 1, pp. 17–38, 2007.

[27] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. of ICASSP*, 2004, vol. 1, pp. 37–40.

[28] P. Rose, T. Osanai, and Y. Kinoshita, "Strength of forensic speaker identification evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian likelihood ratio as threshold," *International Journal of Speech Language and the Law*, vol. 10, no. 2, pp. 179–202, 2003.

[29] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[30] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science*, vol. 309, no. 5736, pp. 892–895, 2005.

[31] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, pp. 193–203, 2000.

[32] J.P. French and P.T. Harrison, "Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases," *International Journal of Speech Language and the Law*, vol. 14, no. 1, pp. 137–144, 2007.

[33] P. Rose and Morrison G.S., "A response to the uk position statement on forensic speaker comparison," *International Journal of Speech Language and the Law*, vol. 16, 2009, In Press.

[34] B. Robertson and G. A. Vignaux, *Interpreting Evidence-Evaluating Forensic Science in the Courtroom*, Wiley, UK, 1995.

[35] R. Cook, I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert, "A hierarchy of propositions: deciding which level to address in casework," *Science and Justice*, vol. 38, no. 4, pp. 231–239, 1998.

[36] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.

[37] J. G. Darboux, P. E Appell, and J. H. Poincaré, *Examen critique des diverses systèmes ou études graphologiques auxquels a donné lieu le bordereau*, pp. 499–600, Ligue franaise de droits de l'homme et du citoyen, Paris, 1908, In French.

[38] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, chapter 1-2, Wiley, 2001.

[39] B. S. Everitt and D. Graham, *Applied Multivariate data Analysis*, ARNOLD, 2001.

[40] C. G. G. Aitken, G. Zadora, and D. Lucy, "A two-level model for evidence evaluation," *Journal of Forensic Sciences*, vol. 52, no. 2, pp. 412–419(8), 2007.

[41] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of decision task performance," in *Proc. of Eurospeech*, 1997, pp. 1895–1898.

[42] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[43] R. K. Ahuja and J. B. Orlin, "A fast scaling algorithm for minimizing separable convex functions subject to chain constraints," *Operations Research*, vol. 49, pp. 784–789, 2001.

[44] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Knowledge Discovery and Data Mining*, 2002.

[45] C. F. Tippett, V. J. Emerson, M. J. Fereday, F. Lawton, A. Richardson, L. T. Jones, and S. M. Lampert, "The evidential value of the comparison of paint flakes from sources other than vehicles," *Journal of the Forensic Science Society*, vol. 2, pp. 61–65, 1968.

[46] G. Strang, "The discrete cosine transform," *SIAM Review*, vol. 41, Number 1, pp. 135–147, 1999.

[47] G. S. Morrison, *Matlab Implementation of Aitken & Lucy's (2004) Forensic Likelihood-Ratio Software Using Multivariate-Kernel-Density Estimation [software]*, 2007, Available: http://geoff-morrison.net.

[48] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in transparent and testable forensic speaker recognition," in *Proc. of Odyssey*, 2006.

[49] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2001.

[50] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the Mixer corpora-2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.

A

Publicación en Interspeech 2009

Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking

Alberto de Castro, Daniel Ramos and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group, Universidad Autonoma de Madrid, Spain.
{a.castro,daniel.ramos,joaquin.gonzalez}@uam.es

Abstract

In this paper we compare forensic speaker recognition with traditional features using two different formant tracking strategies: one performed automatically and one semi-automatic performed by human experts. The main contribution of the work is the use of an automatic method for formant tracking, which allows a much faster recognition process and the use of a much higher amount of data for modelling background population, calibration, etc. This is especially important in likelihood-ratio-based forensic speaker recognition, where the variation of features among a population of speakers must be modelled in a statistically robust way. Experiments show that, although recognition using the human-in-the-loop approach is better than using the automatic scheme, the performance of the latter is also acceptable. Moreover, we present a novel feature selection method which allows the analysis of which feature of each formant has a greater contribution to the discriminating power of the whole recognition process, which can be used by the expert in order to decide which features in the available speech material are important.

Index Terms: automatic formant tracking, forensic speaker recognition, traditional features, likelihood ratio.

1. Introduction

Forensic speaker recognition by human experts using traditional features has being increasingly important in forensic science [1], as more resources have been available to phoneticians in the form of databases and software tools. Despite such progress, the semi-automatic process for generating a result of a forensic comparison is time-demanding in general. In a typical analysis of phonetic-acoustic features, the expert has to perform several steps before a result is obtained [1]. First, the units of interest (words, diphthongs, phonemes, etc.) should be identified and accurately segmented. Second, the phonetic-acoustic features should be extracted from those units, e.g. formant frequencies and formant trajectories. Finally, with those features a comparison should be performed. This whole process may spend a considerable amount of time, as most of the tasks involved are performed manually. This problem becomes more important when the comparison process implies modelling the distribution of features among a relevant population of many speakers, as it happens in likelihood-ratio-based forensic speaker recognition, which is considered a proper way of reporting results to a court [2].

The contribution of this paper is the use of an automatic formant tracking scheme for the use of traditional features in forensic speaker recognition. This allows a much faster recognition process, and therefore, a much higher amount of data can be used as a background set for population modelling, calibration, etc. This also helps to increase the robustness and accuracy of the evidence evaluation process and the validation results from a forensic case. Thus, if an accurate segmentation of the relevant units in the speech signal is available, the rest of the proposed recognition process is automatic. In this work we have used diphthongs segmented by a human experts for comparison, but this labels could also be obtained with a speech recognition system, which would lead to a fully automatic approach of forensic speaker recognition using traditional features. The evaluation of the impact of the proposed recognition scheme with respect to an expert-based semi-automatic formant extraction method is also presented. Moreover, an analysis based on feature selection is performed with the objective of identifying the most discriminative features in the identity inference process. In order to obtain results, a likelihood ratio (LR) approach is used [3, 2]. The human expert performance is taken from the work in [4], whose database is used and experimental set-up replicated. Performance evaluation is given in terms of DET plots and measures of LR performance such as C_{llr} [5].

The paper is organized as follows. In Section 2, the feature extraction approach followed by the expert in [4] will be described. In Section 3 the automatic formant tracking tool used in the paper, developed by [6], is sketched, and the proposed feature selection analysis is detailed. Experiments are presented in Section 4, where the adequacy of the methods proposed with respect to expert-based approaches is shown. Finally, conclusions are drawn in Section 5.

2. Expert-based traditional forensic speaker recognition

In this section we describe the expert-based approach for forensic speaker recognition, which is replicated from [4].

2.1. Database description

The database used in this paper includes recordings of the speech of the 27 male speakers of Australian English from a corpus described in [7] and used in [4]. Sentences are of the kind "Bide, B-I-D-E spells bide". Such utterances contained the target diphthongs which will be used for recognition: /aɪ/, /eɪ/, /oʊ/, /aʊ/ and /ɔɪ/. Their segmentation was performed manually by the human expert by inspection of the spectrogram.

This work has been funded by the Spanish Ministry of Education under project TEC2006-13170-C02-01. We strongly thank Daniel Rudoy from Harvard University for providing the automatic formant tracking software used in this paper; Yuko Kinoshita from the University of Canberra for providing the database for the experiments in this work; and Geoffrey Stuart Morrison from the Australian National University for providing the manual segmentation and the human-supervised semi-automatic formant tracks, and for advice and inspiration.

The speech was recorded with the same microphone in the same environment, and therefore it is not in real forensic conditions, since variability and mismatch are reduced. However, it is a valuable corpus for comparison between automatic and human approaches, since there is a lack of databases segmented and analyzed by human experts.

2.2. Human-in-the-loop feature extraction

For each diphthong manually selected from the database, the formant tracking procedure described in [4] was applied to the first three formants (namely F1, F2 and F3). Once the formant trajectories have been determined, features are extracted by a parametric-curve fitting of the formant trajectories, either polynomial or based on the Discrete Cosine Transform (DCT). As a result, for each formant a variable number of coefficients is selected depending polynomial degree (Equation 1a) or the amount of components in the DCT (Equation 1b):

$$ax^3 + bx^2 + cx + d = 0 \rightarrow (a, b, c, d) \tag{1a}$$

$$X_c(k) = \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos(\frac{k2\pi n}{N}) \rightarrow \\ \rightarrow (X_c(0), X_c(1), X_c(2), X_c(3)) \tag{1b}$$

Thus, for each diphthong analyzed, the feature vector will be formed by the concatenation of the coefficients of the polynomial (e.g., $[a, b, c, d]$) or DCT fitting (e.g., $[X_c(0), X_c(1), X_c(2), X_c(3)]$) for the selected formants. Performance is improved in [4] with equalization of the duration of each diphthong and/or logarithmic frequency scaling applied prior to feature extraction.

2.3. Comparison, LR computation, fusion and calibration

In order to perform a comparison among coefficients, the Multivariate Likelihood Ratio (MVLR) method has been used [3]:

$$LR = \frac{p(\mathbf{x}, \mathbf{y}|\, \theta_p, I)}{p(\mathbf{x}, \mathbf{y}|\, \theta_d, I)} \tag{2}$$

where θ_p is the prosecution hypothesis (*The suspect is the source of the questioned recordings*), θ_d is the defense hypothesis (*Another individual in the relevant population is the source of the questioned recordings*), and $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors to be compared from questioned and control speech material. A function implementing this method in Matlab^{TM} can be found in www.geoff-morrison.net, which we have used in our experiments. See [3] for details.

The comparison strategy is as follows. Every feature vector extracted from a given diphthong found in the questioned speech material (one feature vector for each diphthong occurrence) is compared to all the feature vectors for the same diphthong found in the control material coming from the suspect. Thus, for each comparison, a LR value is computed for each diphthong. Then, the logarithm of the LR values of all the diphthongs are summed (fused) for each comparison in order to improve system performance. Finally, a jackknife linear logistic regression calibration process is applied to the obtained log-LR set as described in [4]. This further calibration procedure of the final, summed log-LR is necessary, since the sum of log-LR values coming from independent sources (e.g., different diphthongs) may not be probabilistically interpretable. This last LR value after calibration will represent the weight of the evidence.

2.4. Performance measures

The determination of the goodness of the LR value computed is achieved by the use of the C_{llr} metric [5]:

$$C_{llr} = \frac{1}{2 \cdot N_p}\sum_{i_p} \log_2\left(1 + \frac{1}{LR_i}\right) + \frac{1}{2 \cdot N_d}\sum_{j_d} \log_2\left(1 + LR_j\right) \tag{3}$$

where N_p and N_d are the number of comparisons (LR values) where θ_p and θ_d are respectively true in the experimental set, also known as target and non-target comparisons. As it can be seen, C_{llr} is an average measure of performance over a given experimental set of LR values, and the higher its value the worse the given LR set.

The overall loss of performance given by C_{llr} can be decomposed into a loss due to discriminating power and another loss due to calibration [5]. In order to test the discriminating power of the proposed methods alone (separation of target and non-target comparisons regardless of the range of the LR values), DET curves are used in automatic speaker recognition. Moreover, C_{llr}^{min} has been also proposed as the optimization of C_{llr} restricted to preserve the discriminating power of the experimental set [5]. Thus, C_{llr}^{min} summarizes a DET curve with a single value, and the calibration of the experimental set is determined by $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$.

3. Traditional forensic speaker recognition using automatic formant tracking

In order to compare the approaches presented in this paper with respect to the one presented in [4], we have replicated the same method and experimental set-up as described in Section 2 with the use of the segmentation labels provided by the human expert, with some differences. First, the semi-automatic formant tracking procedure described in 2.2 has been replaced by a fully automatic process described below [6]. Second, the feature extraction strategies in [4] have been extended with two alternatives which improve system performance. Third, a feature selection algorithm is proposed in order to identify the most discriminant features with a phonetic-acoustic interpretation, which would aid the expert in the selection of the relevant features from the available speech.

3.1. Automatic formant tracking procedure

In order to automatically extract formant trajectories from the speech spectrum for each speech unit (diphthong), the formant tracking tool described in [6] was used. Figure 3.1 shows an example of using this technique. The approach is based on estimating the formant frequencies by means of a Gauss-Markov process. After a cepstral linear prediction analysis, the distribution $p(x_t|y_{1:t})$ is computed for the formant frequencies conditioned to previous waveform data observed, where t is the current time frame, x_t is a state vector result of parameterizing the spectral envelope at time frame t, and $y_{1:t}$ is a function related to the past linear prediction coefficients. Details can be found in [6]. For this work, the formant tracking software was provided by the authors in [6].

3.2. Feature extraction and comparison strategies

For each diphthong, the feature set may vary depending on the fitting (polynomial, DCT) and the time (equalized, not equalized) and frequency (Hz, log-Hz) transformations. In this work we have explored three different feature extraction strategies:

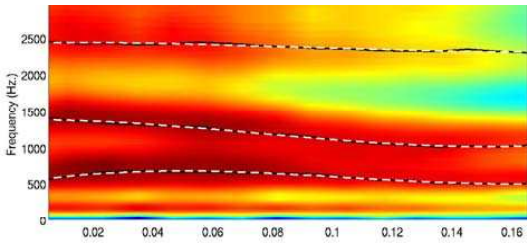


Figure 1: Example of polynomial fitting (degree 3) of /ou/ formant trajectories. Solid black lines are the estimated formant trajectories. Dashed white lines are fitted curves.

- BEST_IND_AUTO: feature set which obtained the best C_{llr}^{min} for each individual diphthong, which gives a different selection for each diphthong (Table 1).

Dipht.	Formants	Fit	f Scale	t Scale
/aɪ/	F1 F2 F3	Poly 3	Hz	Equalized
/eɪ/	F1 F2 F3	DCT 3	Hz	Equalized
/oʊ/	F1 F2 F3	Poly 3	Hz	Equalized
/aʊ/	F1 F2 F3	Poly 3	Hz	Original
/ɔɪ/	F1 F2 F3	Poly 3	Hz	Equalized

Table 1: BEST_IND_AUTO feature extraction scheme.

- BEST_ALL_AUTO: same feature set for all diphthongs, which obtained the best average C_{llr}^{min} value across diphthongs. This strategy encourages a feature set which is more general for all types of diphthongs, being a more reasonable choice if a speech unit not analyzed before is selected for comparison due to limitations in the speech material. The feature extraction selected in this way considered polynomial fitting of degree 3 obtained form F1, F2 and F3 trajectories, natural frequency scale, and equalized duration.
- HUMAN_SEMI: feature set selected by the expert in [4] with semi-automatic human-in-the-loop formant tracking. The feature set is summarized in Table 2.
- HUMAN_AUTO: same feature set as HUMAN_SEMI (Table 2). The main objective of this strategy is the direct comparison of the automatic and the human-in-the-loop formant tracking procedures.

Dipht.	Formants	Fit	f Scale	t Scale
/aɪ/	F1 F2 F3	Poly 3	Hz	Equalized
/eɪ/	F2 F3	DCT 3	Hz	Original
/oʊ/	F1 F2 F3	Poly 3	Hz	Equalized
/aʊ/	F1 F2 F3	Poly 2	Hz	Original
/ɔɪ/	F1 F2 F3	DCT 3	Hz	Original

Table 2: HUMAN_SEMI and HUMAN_AUTO feature extraction schemes.

3.3. Analysis based on feature selection

A feature selection scheme is proposed in order to get a deeper analysis of which specific information in the formants is discriminating. The feature selection algorithm is based on the following steps:

1. For each diphthong and feature in the original feature set, a univariate log-LR set from comparisons in the database is computed, and its C_{llr}^{min} determined. This shows which feature from which formant has a better discriminating power (lowest C_{llr}^{min})
2. The log-LR set from the next feature with lower C_{llr}^{min} value is fused with the output log-LR and if it decreases C_{llr}^{min} value the feature is selected, otherwise the feature is not selected and the sum fusion is undone.
3. The previous step is repeated for all the features in increasing C_{llr}^{min} order.

4. Experiments

4.1. Results on automatic formant tracking

The experiments in this section aim at illustrating the loss of performance due to an automatic approach for formant tracking (HUMAN_AUTO) with respect to a human-in-the-loop semi-automatic formant tracking (HUMAN_SEMI). In table 3 the performance for each diphthong is shown for both strategies. It can be seen that performance in terms of discriminating power (C_{llr}^{min}) of the HUMAN_AUTO approach is worse than for HUMAN_SEMI. However, the HUMAN_AUTO strategy is still acceptable in terms of performance, especially considering that eliminates the need of a human expert for semi-automatic formant selection, consequently reducing the time for a comparison. Figure 2 shows the per-diphthong discriminating power of the HUMAN_AUTO strategy in terms of DET pots.

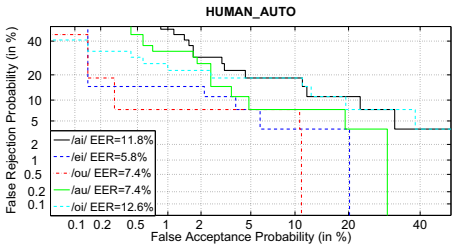


Figure 2: DET plots showing discriminating power for each diphthong with HUMAN_AUTO strategy.

Dipht	HUMAN_SEMI	HUMAN_AUTO
/aɪ/	0.061	0.176
/eɪ/	0.063	0.105
/oʊ/	0.077	0.100
/aʊ/	0.105	0.213
/ɔɪ/	0.082	0.293

Table 3: C_{llr}^{min} values showing discriminating power for each diphthong with HUMAN_AUTO strategy.

In Table 4 the results of fusing and post-calibrating the log-LR values of all the diphthongs for each comparison is shown, both as an overall performance measure C_{llr} and with C_{llr}^{min} as a measure of discriminating power. First, we observe that C_{llr} and C_{llr}^{min} values are quite close for all cases, which indicates a good calibration performance after jackknife logistic regression. This is normal, since jackknife over the test database implies highly matching conditions for calibration. Second, the HUMAN_SEMI strategy [4] achieves better performance than the rest of approaches (in fact, it gives perfect separation $C_{llr}^{min} = 0$). Moreover, although C_{llr} relatively doubles for the best automatic formant tracking procedure, its value remains low in absolute terms (e.g., an increase

of 0.054 from HUMAN_SEMI to BEST_IND_AUTO). Finally, the BEST_ALL_AUTO strategy performs only slightly worse than the BEST_IND_AUTO in absolute terms, which justifies the use of the same feature set for all diphthongs.

	Before selection		After selection	
Strategy	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
BEST_IND_AUTO	0.045	0.110	0	0.0192
BEST_ALL_AUTO	0.074	0.127	0.0058	0.0273
HUMAN_AUTO	0.105	0.181	0.0074	0.0225
HUMAN_SEMI [4]	0	0.056	-	-

Table 4: Performance of automatic formant tracking strategies before and after feature selection.

4.2. Results on feature selection analysis

Table 4 shows C_{llr}^{min} and C_{llr} performance values for the three strategies using automatic formant tracking after the feature selection algorithm proposed in Section 2.2. It can be seen that BEST_IND_AUTO strategy outperforms the rest after feature selection, reaching perfect separation ($C_{llr}^{min} = 0$), and being also better than HUMAN_SEMI before feature selection. Moreover, C_{llr} values for BEST_ALL_AUTO and HUMAN_AUTO after feature selection are also extremely low, indicating excellent performance.

It is worth noting that, due to the low number of comparisons allowed by the database used, the feature selection strategy is applied over the same data in which it is tested. Thus, it is not possible to check if the feature selection strategy improves the performance on new, unseen data. However, this analysis allows to highlight the influence of each formant in the discriminating power of the recognition process. Figure 3 shows a chart representing the final selection of features for the three proposed strategies. It can be seen that, for all cases, F2 seems to contribute with more features to the final selected set, whereas features from F3 are almost not selected, although typically F3 is assumed as significantly discriminating. This is because of the difficulty of reaching a highly accurate automatic extraction of the F3 trajectories. It is worth noting that in the semi-automatic formant tracking procedure followed by HUMAN_SEMI, the final trajectory for F3 is manually chosen among 8 different strategies [4]. This implies a much higher accuracy in F3 formant trajectories for HUMAN_SEMI than for the rest of automatic approaches. Moreover, it can be also seen that for the proposed algorithm /aɪ/ and /oʊ/ are the most feature-contributing diphthongs. These kind of studies may help the expert to decide which units and features are important from the available speech material.

5. Conclusions

In this paper we have presented a comparison among the performance of semi-automatic and automatic formant tracking approaches in forensic speaker recognition using traditional features. The automation of the formant tracking procedure makes the recognition process much faster. Therefore, for each comparison much more data can be used for comparison, which is especially necessary for robust modelling of a relevant population of many speakers in likelihood-ratio based forensic speaker recognition. Results show that performance with automatic formant tracking is worse, but still acceptable. Moreover, we have proposed a feature selection algorithm, which allowed us to analyze the impact of each traditional feature extracted in the discriminating power of the recognition process. Finally, it is

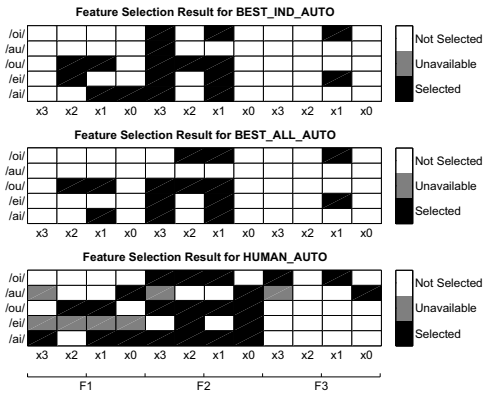


Figure 3: Features selected for automatic formant tracking strategies. Rows are diphthongs. Columns are different degrees of the polynomial or frequency index of the DCT for each formant.

worth noting that, although the database used is small, limited and controlled, expert analysis of a database for forensic speaker comparison is a highly demanding and time consuming process, which requires language proficiency. Thus, such corpora are extremely valuable and rare. Future work is mainly focused on the use of a speech recognizer for diphthong or phoneme segmentation, which would lead to a fully automatic approach for forensic speaker recognition using traditional features. We also plan to test the comparison of automatic and expert-based procedures in more realistic scenarios in terms of speech variability and number of speakers.

6. References

[1] P. Rose, *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, 2002.

[2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[3] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.

[4] G. S. Morrison, "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories," *Journal of the Acoustical Society of America*, vol. 125, no. 4, April 2009, In press.

[5] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[6] D. Rudoy, D. N. Spendley, and P. J. Wolfe, "Conditionally linear Gaussian models for estimating vocal tract resonances," in *Proc. of Interspeech*, Antwerp, Belgium, 2007, pp. 526–529.

[7] Y. Kinoshita and T. Osanai, "Within speaker variation in diphthongal dynamics: What can we compare?," *Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand*, 2006.

B

Presupuesto

1) Ejecución Material

▪ Compra de ordenador personal (Software incluido)	2.000 €
▪ Alquiler de impresora láser durante 6 meses	260 €
▪ Material de oficina	150 €
▪ Total de ejecución material	2.400 €

2) Gastos generales

▪ sobre Ejecución Material	352 €

3) Beneficio Industrial

▪ sobre Ejecución Material	132 €

4) Honorarios Proyecto

▪ 1800 horas a 15 €/ hora	27000 €

5) Material fungible

▪ Gastos de impresión	280 €
▪ Encuadernación	200 €

6) Subtotal del presupuesto

▪ Subtotal Presupuesto	32.774 €

7) I.V.A. aplicable

▪ 16 % Subtotal Presupuesto	5.243,8 €

8) Total presupuesto

▪ Total Presupuesto	38.017,8 €

Madrid, Octubre 2009
El Ingeniero Jefe de Proyecto

Fdo.: Alberto de Castro Rodríguez
Ingeniero Superior de Telecomunicación

C

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un *Desarrollo de un sistema de reconocimiento forense de locutor utilizando parámetros fonético-acústicos y extracción automática de formantes, y comparación con extracción manual por parte de expertos.*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales.

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares.

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.