UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR
DEPARTAMENTO DE TECNOLOGÍA
ELECTRÓNICA Y DE LAS COMUNICACIONES

RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR E IDIOMA MEDIANTE CARACTERIZACIÓN ACÚSTICA DE UNIDADES LINGÜÍSTICAS

–TESIS DOCTORAL–

Autor: Javier Franco Pedroso
(Ingeniero de Telecomunicación,
Universidad Politécnica de Madrid)

Madrid, Mayo de 2016

Colophon

This book was typeset by the author using LaTeX2e. The main body of the text was set using a 11-points Computer Modern Roman font. All graphics and images were included formatted as Encapsuled Postscript (™ Adobe Systems Incorporated). The final postscript output was converted to Portable Document Format (PDF) and printed.

Departamento:	Tecnología Electrónica y de las Comunicaciones Escuela Politécnica Superior Universidad Autónoma de Madrid (UAM) ESPAÑA
Tesis:	Reconocimiento automático de locutor e idioma mediante caracterización acústica de unidades lingüísticas
Autor:	**Javier Franco Pedroso** Ingeniero de Telecomunicación (Universidad Politécnica de Madrid)
Director:	**Joaquín González Rodríguez** Catedrático de Universidad Universidad Autónoma de Madrid
Año:	2016
Tribunal:	Presidente: **Luis Hernández Gómez** Universidad Politécnica de Madrid
	Secretario: **Doroteo Torre Toledano** Universidad Autónoma de Madrid
	Vocal 1: **Juana Gil Fernández** Consejo Superior de Investigaciones Científicas (CSIC)
	Vocal 2: **Ascensión Gallardo Antolín** Universidad Carlos III de Madrid
	Vocal 3: **Javier Hernando Pericás** Universidad Politécnica de Cataluña

Las investigaciones descritas en esta Tesis se han llevado a cabo en el Grupo de Reconocimiento Biométrico ATVS, Departamento de Ingeniería Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid. Dichas investigaciones han sido parcialmente financiadas por el Ministerio de Economía y Competitividad a través del proyecto MICINN TEC2012-37585-C02-01, por el Ministerio de Ciencia e Innovación a través del proyecto TEC2009-14719-C02-01 y por la Cátedra Universidad Autónoma de Madrid-Telefónica.

Los trabajos preliminares de esta Tesis se recogen en el Trabajo Fin de Máster del autor, "Temporal Contours in Linguistic Units for Automatic Text-Independent Speaker Recognition", que fue finalista del "Premio MAVIR 2013 al Mejor Trabajo de Máster en Tecnologías de la Lengua aplicadas a los Sistemas Inteligentes de Acceso a la Información Multilingüe y Multimedia".

Parte de los trabajos presentados en esta Disertación se recogen en la siguiente publicación, que fue premiada con el "Microsoft Research Best Student Paper Award" *en la conferencia internacional* SIG-IL/Microsoft 2009*:"Multilevel and Channel-compensated Language Recognition: ATVS-UAM systems at NIST LRE 2009", Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano e Joaquin Gonzalez-Rodriguez, Proc. of SIG-IL/Microsoft 2009, Porto Salvo, Portugal, September 2009.*

Derivado de parte de los trabajos presentados en esta Disertación, el autor fue invitado a dar sendas ponencias en las "VI Jornadas MAVIR: Tecnologías de Acceso a la Información: Estado actual y retos", Madrid (España) *y en el* "15th Meeting of the ENFSI (European Network of Forensic Science Institute) Working Group for Forensic Speech and Audio Analysis (FSAAWG), Helsinki (Finlandia)".

Resumen

El reconocimiento automático de personas mediante la voz, o reconocimiento automático de locutores, ha experimentado en los últimos años un gran avance gracias a la incorporación de técnicas de caracterización y modelado cada vez más complejas, situando a los sistemas biométricos basados en este rasgo en tasas de error impensables hace una década. Si bien estas técnicas han permitido mejorar la seguridad biométrica, entendida como sistemas de control de acceso o aquellos que permiten evitar la suplantación de la identidad de las personas, aspectos importantes como el carácter interpretable de dichas características y modelos se han dejado de lado. La misma situación se da en el ámbito del reconocimiento de idioma, donde el objetivo de mejorar las tasas de identificación está muy por encima del de obtener otros tipos de información que permitan interpretar qué distingue a una lengua de otra. En este contexto, esta Tesis se centra en el uso y aplicación de técnicas de caracterización y modelado con carácter eminentemente interpretable, de forma que permitan no sólo distinguir entre locutores o idiomas, sino también ofrecer información de qué aspectos son los que los hacen diferentes.

Una de las herramientas empleadas en esta Tesis para este fin es el uso de relaciones de verosimilitud. La relación de verosimilitud (o likelihood ratio, LR) se ha adoptado en el ámbito forense como el marco teórico más apropiado para la presentación de evidencias, y representa el cociente de probabilidades de la evidencia dadas dos hipótesis opuestas: que las muestras dubitadas e indubitadas proceden de la misma fuente, por una parte, y que proceden de fuentes diferentes, por otra. Así, el resultado de la comparación de dos conjuntos de muestras tiene una interpretación directa, frente a las puntuaciones crudas, carentes de significado por sí mismas, proporcionadas por los sistemas de identificación clásicos. Pero además, facilita la combinación de información procedente de diversas fuentes gracias a su interpretación probabilística.

Esta Disertación comienza afrontando el problema de la segmentación de audio, una etapa crucial para la identificación de locutor e idioma en entornos no controlados ya que permite aislar los tramos de la grabación que sólo contengan voz. La segmentación de audio tiene por objetivo general dividir el flujo de audio en segmentos homogéneos en lo que al contenido acústico se refiere. Esta segmentación suele abordarse como un problema de clasificación, en el que cada trama de audio debe asignarse a una de las posibles clases (por ejemplo, voz con ruido de fondo, voz con música, voz aislada, etc.). Dada la amplitud de esta casuística, en esta Tesis se plantea la segmentación como un problema de detección de clases acústicas en un sentido más amplio, de forma que la segmentación final venga dada por la combinación de varios de estos detectores. Así, por ejemplo, se tendría un detector de voz en cualquier contexto (voz asilada, voz con ruido de fondo, voz con música de fondo, etc.), un detector de ruido (ruido asilado, ruido en presencia de voz, etc.) y un detector de música (música aislada, etc.). Cada uno de estos detectores asigna un LR a cada trama de audio, de forma que sus salidas pueden combinarse de forma natural para determinar la presencia de alguna de las posibles combinaciones de ellos, e incluso la ausencia

de todos ellos.

La forma habitual de obtener LRs en el ámbito del reconocimiento automático consiste en aplicar una transformación a las puntuaciones crudas proporcionadas por el sistema. Para ello, es necesario disponer de un conjunto de datos adicional a partir del cual entrenar dicha transformación. Sin embargo, en el ámbito forense los datos suelen ser un recurso escaso, por lo que la aproximación habitual está basada en modelos probabilísticos que dan lugar a la obtención de LRs directamente a partir de las características empleadas. El modelo concreto a aplicar depende de la distribución de muestras en la llamada población de referencia; cuando esta distribución no puede aproximarse de forma paramétrica, debe recurrirse a técnicas de estimación para caracterizarla. En esta Tesis también se aborda este problema, proponiéndose una nueva aproximación mediante mezclas de modelos de gaussianas (gaussian mixture models, GMMs) frente a la clásica estimación mediante una función kernel.

Otra de las herramientas empleadas en esta Tesis para proporcionar información interpretable adicional al proceso de identificación en sí mismo es la caracterización de locutor e idioma en términos fonético-acústicos. En el caso del reconocimiento de idioma, se ha combinado la información procedente de sistemas acústicos clásicos (basados en características cepstrales) con la proporcionada por sistemas fonotácticos. Estos sistemas hacen uso de la información fonética para construir un modelo del idioma en base a la frecuencia de repetición de determinadas secuencias de fonemas (n-gramas), con la particularidad de que la fonética del reconocedor puede ser distinta a la del idioma que se pretende identificar. Como se mostrará, la combinación de varios sistemas fonotácticos junto con sistemas acústicos da lugar a mejoras significativas, además de proporcionar información fácilmente interpretable sobre las diferencias de unos idiomas respecto a otros.

Una aproximación similar se ha seguido para abordar el problema del reconocimiento de locutores. Una de las aplicaciones más directas de los sistemas de reconocimiento automático de locutores es el reconocimiento de locutor forense, donde intenta determinarse si la persona que habla en una grabación determinada (muestra dubitada) es el acusado en cuestión, a partir otras grabaciones de éste (muestras indubitadas). Sin embargo, el reconocimiento de locutor forense y el reconocimiento automático de locutor han seguido caminos separados debido, en gran parte, a la dificultad de interpretar los resultados obtenidos por estos últimos, que son vistos por la mayoría de la comunidad forense como sistemas de caja negra.

Esto es debido, por una parte, a que los sistemas automáticos se basan en características cuya relación con las propiedades anatómicas de los individuos se desdibujan debido a la cadena de procesado a la que se somete a la señal de voz con el objetivo de eliminar la componente de señal indeseada y de realzar la información discriminante. Y por otra, a que los sistemas automáticos reducen la comparativa entre dos grabaciones de voz a una única puntuación que integra toda la información presente en ambas grabaciones. La comparación de voz forense, en cambio, suele hacer uso de características directamente ligadas a aspectos anatómicos del individuo, como por ejemplo las frecuencias formantes. Así mismo, es habitual que la comparación se realice atendiendo a criterios fonético-acústicos, comparando unidades equivalentes entre sí desde un

punto de vista lingüístico.

En esta Tesis se aborda el problema del reconocimiento de locutor desde una perspectiva que intenta ligar ambas ramas, automática y forense. Para ello, se hace uso de sistemas automáticos para segmentar la señal de audio en base al contenido fonético y extraer características fácilmente interpretables como las frecuencias formantes. A partir de esta información, se construyen sistemas automáticos de reconocimiento de locutor independientes para cada unidad lingüística, lo que permite analizar qué unidades son más discriminativas en término medio, o si los locutores presentan particularidades que les hacen más distinguibles entre sí en base a determinadas unidades lingüísticas. Además, la información discriminativa repartida entre las distintas unidades lingüísticas puede combinarse de forma natural gracias la obtención de relaciones de verosimilitud para cada unidad.

Abstract

Automatic recognition of speakers from their voices, or automatic speaker recognition, has experienced a great advance in recent years with the addition of feature extraction and modeling techniques increasingly complex, which have boosted biometric systems based on this trait up to error rates unthinkable a decade ago. Even though these techniques have allowed improving biometric security, in the sense of access control systems or those that avoid the impersonation of people, important issues as the interpretable nature of these features and models have been left apart. The same situation happens in the language identification field, in which the goal of improve the identification rates is much more important than obtaining other types of information that allow to interpret what makes a language different from another. In this context, this Thesis is focused on the use and application of feature extraction and modeling techniques eminently interpretable, allowing not only to distinguish between speakers or languages, but also to provide information regarding which aspects makes them different.

One of the tools used in this Thesis to achieve this goal is the use of likelihood ratios. The likelihood ratio (LR) has been adopted in the forensic field as the more convenient theoretical framework for evidence reporting, and is the ratio between the likelihoods of the evidence under two competing hypotheses: that both control and recovered samples come from the same source, on the one hand, and that control and recovered samples come from different sources, on the other hand. Thus, the result of the comparison of two sets of samples has a straightforward interpretation, unlike the raw scores provided by classical identification systems, meaningless by themselves. Moreover, the use of LRs eases the combination of different information sources thanks to its probabilistic interpretation.

This Dissertation begins by tackling the problem of audio segmentation, a crucial stage for speaker and language identification in uncontrolled scenarios, as it allows isolating only-speech segments in audio recordings. The general goal of audio segmentation is to split the audio stream in homogeneous segments regarding the acoustic content. Audio segmentation is usually tackled as a classification problem in which each audio frame must be assigned to one of the possible classes (for instance, speech with noise in the background, speech with music in the background, isolated speech, etc.). Due to the large number of possible combinations, audio segmentation is tackled in this Thesis as a detection problem of acoustic classes defined in a broader sense, being the final segmentation given by the combination of several of these detectors. Thus, the proposed segmentation system consists in a speech detector for speech in any given context (isolated speech, speech with noise in the background, speech with music in the background, etc.), a noise detector (isolated noise, noise with speech in the foreground, etc.) and a music detector (isolated music, etc.). Each of these detectors provides a LR for each audio frame so that their outputs can be naturally combined to determine the presence of any of the possible combinations, or even the absence of all of them.

Likelihood ratios are usually obtained in the automatic recognition field by transforming the raw scores provided by the system. In order to do that, an additional dataset is needed in order to train the transformation step. However, data is scarce in forensics, so the usual approach is based on probabilistic models that directly provide LRs from the features. The specific model to be applied depends on the distribution of samples in the so-called reference population; when this distribution cannot be approximated in a parametric way, density estimation techniques are required in order to model it. This problem is also faced in this Thesis, and a new approach is proposed that uses Gaussian mixture models (GMMs) instead of the classical estimation through kernel functions.

Another tool used in this Thesis to provide interpretable information in addition to the identification process itself is the speaker and language characterization through acoustic-phonetic information. In the case of language recognition, the information provided by classical acoustic systems (based on cepstral features) have been combined with that provided by phonotactic systems. Phonotactic systems make use of phonetic information in order to train a language model based on the frequency of occurrence of particular phoneme sequences (n-grams), with the particularity that the language of the phone recognizer may be different to the language we are attempting to identify. As it will be shown, the combination of several phonotactic systems with acoustic systems gives rise to significant improvements, in addition to provide easily interpretable information regarding the differences between languages.

A similar approach has been followed in order to tackle the speaker recognition problem. One of the most straightforward applications of automatic speaker recognition systems is forensic speaker recognition. Forensic speaker recognition attempts to determine if the person speaking in some given recording (recovered sample) is the suspect, by using some additional recordings from him/her (control samples). However, forensic speaker recognition and automatic speaker recognition have gone their separate ways due, to a great extent, to the difficulty of interpret the results obtained by the latter ones, which are seen by most of the forensic community as black-box systems.

This is due, on the one hand, to the fact that automatic systems are based on features whose relationship with anatomical characteristics of individuals is blurred due to the processing chain used to remove the undesired component in the audio signal and to enhance the discriminative information. And, on the other hand, to the fact that automatic systems summarize the comparison between two given speech recordings into one single score integrating all the information present in both recordings. Instead, forensic voice comparison is usually performed using features directly related to anatomical traits, as for example formant frequencies. Likewise, it is usual to perform the comparison according to acoustic-phonetics criteria, comparing equivalent units from a linguistic point of view.

In this Thesis, the speaker recognition problem is tackled from a perspective that attempts to link both branches, automatic and forensic. To do this, automatic systems are used to segment the audio signal based on the phonetic content and to extract easily interpretable features as formant frequencies. Starting with this information, independent automatic speaker recog-

nition systems are developed for each linguistic unit, allowing to analyze which units are more discriminative on average, or whether speakers have particularities that allow better distinguish between them by using specific linguistic units. Furthermore, the discriminative information spread among the different linguistic units can be naturally combined thanks to obtaining likelihood ratios for each unit.

A mis padres y hermanos.

A mi tía Cándida.

A Gema.

A Greta y Eric.

Agradecimientos

En primer lugar, me gustaría dar las gracias a mi director de Tesis, el Profesor Joaquín González, por guiar mis pasos en este largo camino y por la confianza que ha depositado en mí para afrontar éste y otros proyectos. Esta Tesis no hubiera sido posible sin los empujones que me ha hecho dar en el último año y su ayuda con los trámites finales. Así mismo, quiero agradecer el apoyo recibido de los Profesores Javier Ortega y Doroteo Torre.

Son muchos los compañeros con los que he tenido la suerte de convivir desde que recalé en el ATVS allá por 2004, cuando aún no era Ingeniero Técnico siquiera. Vayan por delante aquellos a los que siempre vi como un ejemplo a seguir: Alberto Montero, Julián Fiérrez y Daniel Ramos. Guardo especial afecto a los que, junto a este último, durante mucho tiempo fueron mis más fieles colaboradores, dentro y fuera del laboratorio: Javier González e Ignacio López. Por suerte para mí, el vacío que dejaron en el laboratorio lo ocupan ahora personas como Alicia Lozano y Rubén Zazo.

Otras personas con las que he tenido la suerte de coincidir durante muchos años a lo largo de mi estancia en el ATVS son Javier Galbally, Pedro Tomé, Rubén Vera, Marta Gómez, Ruifang Wang y Ram Prasad. Y, por supuesto, no puedo olvidarme del resto de compañeros actuales, que hacen la vida en el laboratorio realmente agradable: Ester González, Rubén Tolosana y Aythami Morales. Aparte de vosotros, sé que debo agradecer también a un sinfín de compañeros con los que tuve la suerte de compartir laboratorio, pero me temo que mi memoria no dé para todos. Perdonad aquellos a los que olvide; los que me vienen a la mente, sin ningún orden particular, son: Miriam Moreno, Verónica Peña, María Puertas, Almudena Gilperez, Álvaro Diéguez, Ismael Mateos, Daniel Hernández, Alejandro Abejón, Lucas Pérez, Manuel Freire, Fernando Espinoza, Sergio Lucas y Víctor González.

En el plano más personal, debo agradecer a mis amigos de toda la vida, precisamente, que lleven ahí toda la vida, pase lo que pase. A mis hermanos. Por supuesto, a mis padres, pues os debo todo lo bueno que he conseguido en la vida, además de la vida misma. A Gema, Greta y Eric, por llenar mi vida de amor y felicidad. Estas palabras no son más que una burda aproximación a todo lo que siento por vosotros.

Javier Franco Pedroso
Madrid, Mayo de 2016

Índice de contenidos

Lista de figuras

Bloque 1

Introducción

1.1. Motivación

Los sistemas automáticos de reconocimiento de locutor e idioma han experimentado en los últimos años un gran avance gracias al desarrollo de técnicas de reconocimiento de patrones y aprendizaje automático cada vez más complejas. Si bien estas técnicas han permitido reducir las tasas de error de los sistemas de reconocimiento hasta límites impensables hace una década (Gonzalez-Rodriguez [2014]), los procedimientos aplicados, debido al bajo nivel al que operan, dificultan la interpretación de los resultados obtenidos. Es decir, permiten determinar con gran precisión que la voz en dos grabaciones corresponde al mismo (o distinto) locutor/idioma, pero no así explicar porqué (Gonzalez-Rodriguez *et al.* [2014]). Por otra parte, la acústica fonética es una rama de la fonética que estudia la voz desde el punto de vista de sus características acústicas, relacionando la información lingüística con los sonidos del habla codificados en la onda sonora. Así, este análisis proporciona información de alto nivel que, incorporada al proceso de reconocimiento, puede dar lugar a resultados fácilmente interpretables de gran utilidad, especialmente en el ámbito forense. Con este objetivo, esta Tesis se centra en el desarrollo de sistemas automáticos de reconocimiento de locutor e idioma mediante caracterización fonético-acústica.

El campo de la comparación de voz forense podría beneficiarse en gran medida de la adopción de sistemas automáticos por varios motivos. Por una parte, ahorrarían gran cantidad de trabajo humano, ya que el procesado de las grabaciones suele realizarse de forma manual. Por otra parte, al no haber intervención humana, se eliminarían posibles subjetividades en la toma de decisiones, que pueden llevar a conclusiones distintas en condiciones similares; aunque los sistemas automáticos no están libres de error, dichos errores son sistemáticos. Finalmente, los sistemas automáticos son fácilmente evaluables, pudiéndose comprobar tanto su capacidad discriminativa como sus propiedades de calibración en conjuntos de datos tan grandes como se quiera en un tiempo reducido.

A pesar de estas ventajas potenciales, varios estudios (Gold and French [2011]; Morrison *et al.* [2016]) revelan que la mayoría de los laboratorios forenses continúan aplicando procesos manuales o semi-automáticos para la comparación de voz (ver Figura 1.1). Esto es debido en gran medida

a que el reconocimiento automático sigue una aproximación muy diferente a las empleadas tradicionalmente por los expertos forenses. Por ejemplo, en las aproximaciones acústico-fonéticas, la comparación de voz suele centrarse en la comparación de los sonidos entre unidades lingüísticas equivalentes (Rose [2002]). En cambio, los sistemas automáticos en el estado del arte utilizan todo el contenido acústico de ambas grabaciones independientemente del contexto lingüístico (Dehak *et al.* [2011]; Kinnunen and Li [2010]), lo que les lleva a ser vistos en este ámbito como sistemas de *caja negra*. Por lo tanto, el desarrollo de sistemas de reconocimiento automático que implementen aproximaciones tradicionales puede facilitar la adopción de estos sistemas por parte de la comunidad forense.

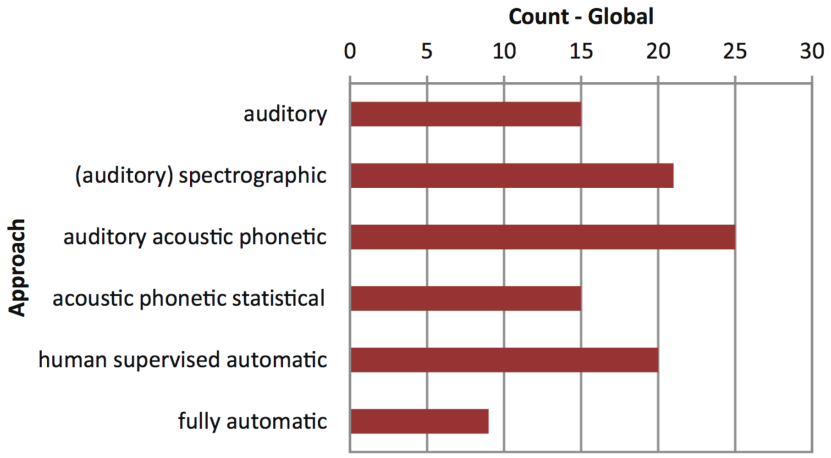


Figura 1.1: *Aproximaciones al reconocimiento forense de locutor y número de países que las utilizan. Fuente: Morrison* et al. *[2016].*

Para que este proceso de adopción pueda llevarse a cabo, es necesario también que la evaluación de evidencias por parte de los sistemas automáticos se ajuste al marco probabilístico de presentación de resultados cada vez más aceptado en ciencias forenses (Figura 1.2). Este marco establece, entre otras cosas (European Network of Forensic Science Institutes [2015]), que la evaluación de evidencias debe realizarse por medio de la asignación de una relación de verosimilitudes (*likelihood ratio*, LR). El LR mide el grado de apoyo que las pruebas proporcionan a las dos proposiciones o hipótesis alternativas de interés: que el acusado está en el origen de las muestras cuestionadas o que se trate de cualquier otro individuo. En el ámbito del reconocimiento automático de locutor (Gonzalez-Rodriguez *et al.* [2007]), los LRs se obtienen por medio de la transformación de las puntuaciones *crudas* proporcionadas por el sistema (*scores*), mediante una etapa adicional conocida como *calibración* (Brümmer and du Preez [2006]; Ramos-Castro *et al.* [2006]). Para obtener LRs por medio de este proceso (conocidos como *score-based LRs*), es necesario un conjunto de datos adicional que permita obtener los parámetros de la transformación. En esta Tesis, se aborda también el problema de la obtención de LRs mediante un método alternativo en el que el proceso de comparación proporciona directamente relaciones de verosimilitud (*feature-based LRs*), evitando así el proceso de calibración posterior que requiere datos adicionales para su entrenamiento.

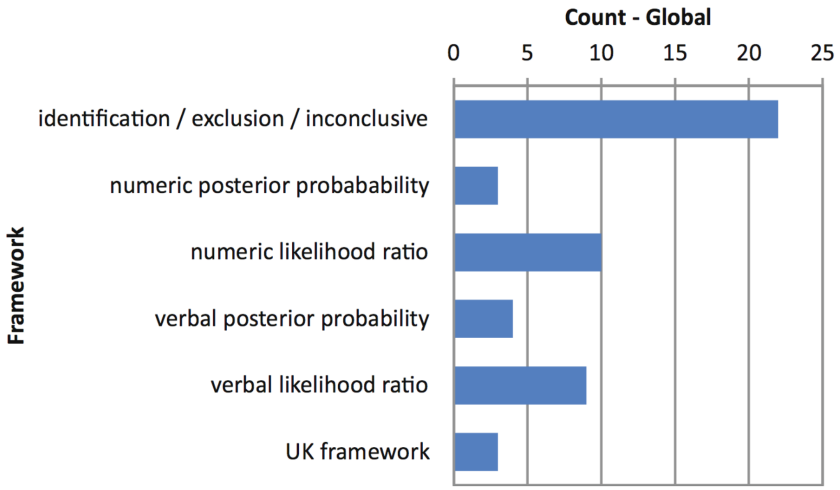


Figura 1.2: *Marcos de evaluación de evidencias y número de países que las utilizan. Fuente: Morrison* et al. *[2016].*

Finalmente, debe tenerse en cuenta que las grabaciones de voz, especialmente las que se presentan en el ámbito forense, pueden adquirirse en condiciones no controladas, apareciendo en ese caso fuentes acústicas distintas a la de interés (voz), como por ejemplo ruido o música, que pueden darse de forma simultánea o no junto con la señal de voz. Por ello, una etapa previa crucial para el reconocimiento de locutor o idioma es la segmentación de audio, ya que permite aislar los tramos de la grabación que contengan voz. Si, además, el sistema de segmentación fuese capaz de proporcionar LRs, podría integrarse en el marco probabilístico de evaluación de evidencias para su aplicación en entornos forenses. En esta Tesis, además, se aprovecha el marco probabilístico que proporcionan los LRs para abordar la segmentación de audio como un problema de detección de clases acústicas mediante detectores calibrados, lo que permite combinarlos de forma flexible.

1.2. Objetivos

Teniendo en cuenta el carácter multidisciplinar de esta Tesis, son varios los objetivos que se persiguen:

- estudiar y validar aproximaciones que combinen técnicas tradicionales (fonético-acústica) y automáticas al reconocimiento del locutor e idioma con el fin de aprovechar sus sinergias. Por una parte, el reconocimiento automático permite el procesado eficiente de grandes cantidades de datos integrando técnicas muy potentes de reconocimiento de patrones, obteniendo así resultados estadísticamente concluyentes y reproducibles. Por otra, las técnicas tradicionales aportan resultados interpretables basados en características directamente relacionadas con las particularidades fisiológicas y anatómicas de los individuos, de gran valor en aplicaciones forenses.

- analizar la aplicabilidad en entornos forenses de las aproximaciones así desarrolladas. La evaluación de evidencias en entornos forenses debe realizarse en forma de relaciones de verosimilitud, para lo cuál los sistemas desarrollados deben proporcionar este tipo de información a su salida, no bastando con las habituales puntuaciones *crudas* de similitud que suelen proporcionar los sistemas de reconocimiento biométrico. Además, se requiere que dichas relaciones de verosimilitud sean útiles, proporcionando información fiable sobre el resultado de la comparación; es decir, deben estar calibradas.

- estudiar técnicas de cálculo de LRs directamente a partir de características y validar su aplicación en sistemas de reconocimiento automático. Existen varias técnicas de cálculo de LRs a partir de características que se han propuesto en otras disciplinas forenses; sin embargo, el procedimiento habitual en sistemas de reconocimiento automático es obtener una puntuación de similitud y luego transformarla a un LR mediante un proceso de calibración. Aunque este procedimiento en dos pasos no resta validez a los LRs así obtenidos, tiene el inconveniente de que el proceso de calibración debe ser entrenado, lo que requiere de datos adicionales.

- estudiar y validar técnicas de segmentación de audio que puedan integrarse en el proceso de toma de decisiones forense. Aunque existen gran cantidad de aproximaciones a la segmentación de audio, estas realizan una asignación *rígida* a una de las posibles clases consideradas, siendo éstas mutuamente excluyentes. La integración de un marco probabilístico en forma de LRs en este proceso puede aportar información útil cuando se emplean estos sistemas automáticos en entornos forenses, por ejemplo considerando la incertidumbre de que las muestras de voz a evaluar incorporen, además, fuentes acústicas de otra naturaleza (ruido o música) que influyan en el resultado de la comparación.

- finalmente, divulgar los resultados obtenidos es un aspecto fundamental para poner en conocimiento de la comunidad científica los estudios realizados. En este sentido, son de especial interés tanto los foros de reconocimiento automático como los de ciencias forenses.

1.3. Trabajos desarrollados: aportación original del autor

En esta sección se enumeran los trabajos desarrollados a lo largo de esta Tesis Doctoral. En primer lugar se detallan las publicaciones compendiadas, aquellos trabajos que constituyen hitos de importancia destacada debido a su publicación como artículos de revista o capítulo de libro. Así mismo, se compendian algunos artículos de congreso que complementan a alguna de estas publicaciones o ligan varias de ellas. En segundo lugar se detallan otros trabajos relacionados con la tesis que han dado lugar a publicaciones de congreso, y que en muchos casos son trabajos previos de las publicaciones compendiadas o desarrollos intermedios.

1.3.1. Publicaciones compendiadas

ATVS-UAM System Description for the Albayzin 2014 Audio Segmentation Evaluation. **Javier Franco-Pedroso**, Elena Gomez Rincon, Daniel Ramos and Joaquin Gonzalez-Rodriguez. Proceedings of IberSPEECH 2014: "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop", November 19-21th, 2014, LasPalmas de Gran Canaria (Spain), pp: 247-252

Este trabajo describe en detalle el sistema de segmentación automática de audio diseñado e implementado para la evaluación competitiva Albayzín 2014 (Ortega *et al.* [2014]) organizada por la Red Temática en Tecnologías del Habla (RTTH). La segmentación de audio se aborda habitualmente mediante una de las dos siguientes aproximaciones: *segmentación basada en distancia*, en la que se evalúa la similitud entre dos conjuntos de datos en ventanas adyacentes para determinar los puntos de cambio entre clases acústicas, y el posterior agrupamiento o clasificación en clases acústicas mutuamente excluyentes de los segmentos así definidos; y *segmentación basada en modelos*, en la que los distintos segmentos surgen a partir de la clasificación directa de las tramas de audio por comparación con modelos predefinidos de clases acústicas mutuamente excluyentes. El sistema planteado por el autor, en cambio, aborda el problema de la segmentación mediante el uso de detectores calibrados de clases acústicas en un sentido amplio, de forma que las clases predefinidas no son mutuamente excluyentes entre sí, sino que pueden estar solapadas. En este sistema, la segmentación final viene dada por la combinación de las relaciones de verosimilitud obtenidas por cada uno de estos detectores.

Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains. Diego Castán, David Tavarez, Paula Lopez-Otero, **Javier Franco-Pedroso**, Héctor Delgado, Eva Navas, Laura Docio-Fernández, Daniel Ramos, Javier Serrano, Alfonso Ortega and Eduardo Lleida. EURASIP Journal on Audio, Speech, and Music Processing. 2015 Dec;2015(1):1-9. doi: 10.1186/s13636-015-0076-3

Factor de impacto[1]: 0.386/0.624 (Q4)

Esta publicación recoge los resultados del sistema de segmentación automática de audio descrito en el anterior trabajo, junto con los del resto de grupos de investigación participantes en la evaluación Albayzín 2014 de segmentación de audio (Ortega *et al.* [2014]). Así mismo, para esta publicación se llevaron a cabo experimentos de fusión entre los sistemas de los distintos participantes. Aunque el rendimiento del sistema diseñado por el autor no estuvo entre los sistemas ganadores, los experimentos de fusión demuestran que la aproximación seguida proporciona información muy complementaria a aquellos, pues la fusión que obtenía mejores resultados era una de las que incluía el sistema diseñado por el autor.

[1]Datos extraídos de Thomson Reuters 2012. Factor de impacto último año (2014)/últimos 5 años (cuartil).

Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. **Javier Franco-Pedroso**, Daniel Ramos and Joaquin Gonzalez-Rodriguez. PLoS ONE, 2016 Feb;11(2):1-25. doi: 10.1371/journal.pone.0149958

Factor de impacto: 3.234/3.702 (Q1)

En el ámbito forense, es común el uso de un marco probabilístico para la obtención de relaciones de verosimilitud en términos bayesianos que permita determinar si dos conjuntos de muestras proceden o no de la misma fuente. Una de las técnicas más empleadas (Aitken and Lucy [2004]; Zadora *et al.* [2014]) hace uso de un modelo generativo para representar las muestras a evaluar en términos de la variabilidad intra-fuente y de la variabilidad inter-fuente. En él, la variabilidad intra-fuente se asume distribuida en forma de gaussiana multivariada mientras que la variabilidad inter-fuente puede no estarlo. Cuando la variabilidad inter-fuente no responde a una distribución paramétrica conocida, debe estimarse su función densidad de probabilidad mediante alguna técnica. Una de las más usadas es la estimación mediante funciones *kernel*. En este trabajo compendiado se exponen las limitaciones de esta técnica y se propone el uso de modelos de mezclas de gaussianas para la estimación de dicha función densidad de probabilidad. Así mismo, se derivan las expresiones necesarias para el cálculo de LRs considerando este tipo de distribución, y se compara su eficacia respecto a las funciones *kernel* en varias bases de datos forenses de distinta naturaleza.

Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009. Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, **Javier Franco-Pedroso**, Daniel Ramos, Doroteo Torre Toledano and Joaquin Gonzalez-Rodriguez. IEEE Journal of Selected Topics in Signal Processing, 2010 Sep. 4(6):1084-1093. doi: 10.1109/JSTSP.2010.2076071

Factor de impacto: 2.373/3.681 (Q1)

Este trabajo presenta el sistema desarrollado para la evaluación de reconocimiento de idioma organizada por el Instituto Nacional de Estándares y Tecnología estadounidense (NIST, por sus siglas en inglés) en el año 2009 (National Institute of Standards and Technology [2009]). Dicho sistema consistía en la combinación de sub-sistemas que explotaban, de forma independiente, información acústica (bajo nivel) y fonética (alto nivel). La información fonética se utilizaba para crear un modelo del idioma en base a la frecuencia de aparición de determinadas secuencias de fonemas, lo que se conoce como sistemas *fonotácticos*. Aunque el uso de sistemas *fonotácticos* ya era habitual en reconocimiento de locutor e idioma, en esta publicación se explota la combinación de un gran número de este tipo de sistemas entre sí, primero, y con sistemas acústicos después, mediante la técnica conocida como *anchor modelling*. Esta técnica permite desarrollar un modelo para un idioma específico aunque no se disponga de audio de entrenamiento suficiente, utilizando para ello las puntuaciones de las grabaciones de test de ese idioma frente a los modelos acústicos

o *fonotácticos* de otros idiomas.

Fine-grained automatic speaker recognition using cepstral trajectories in phone units. **Javier Franco-Pedroso**, Joaquin Gonzalez-Rodriguez, Javier Gonzalez-Dominguez and Daniel Ramos. Quantitative approaches to problems in linguistics. LINCOM Studies in Phonetics 08, 2012. ISBN: 9783862883844

Este trabajo presenta la primera aproximación automática al reconocimiento de locutor forense mediante caracterización de trayectorias cepstrales en unidades lingüísticas. Aunque las trayectorias cepstrales en unidades lingüísticas habían sido empleadas con anterioridad en reconocimiento automático de locutor, únicamente se había empleado la información lingüística en el proceso de extracción de características, prescindiendo luego de esta información a la hora de caracterizar a los distintos locutores. En este trabajo se construyen sistemas automáticos de reconocimiento de locutor independientes para cada fonema, de forma que el proceso de comparación entre dos grabaciones de voz se descompone en función de los fonemas comunes entre ambas locuciones. Esto permite establecer una relación con procedimientos seguidos en otros campos del reconocimiento de locutor forense, como por ejemplo la acústica fonética, poniendo a disposición de la comunidad científica los resultados obtenidos para distintos fonemas. Así mismo, se analizan diversas estrategias de selección de unidades y su fusión para integrar estas informaciones en una única relación de verosimilitud por comparación.

Linguistically-constrained formant-based i-vectors for automatic speaker recognition. **Javier Franco-Pedroso** and Joaquin Gonzalez-Rodriguez. Speech Communication, 2016 Feb;76(C):61-81. doi: 10.1016/j.specom.2015.11.002
Factor de impacto: 1.256/1.786 (Q2)

Siguiendo la línea de la publicación anterior, este trabajo profundiza en las técnicas de caracterización del locutor a partir de información directamente relacionada con las características anatómicas del individuo y actualiza la tecnología de reconocimiento automático empleada en el proceso de verificación. Así, se hace uso de frecuencias formantes y su dinámica temporal en unidades lingüísticas para caracterizar a los locutores, extrayendo un i-vector por unidad lingüística para cada locución. Al igual que en la publicación anterior, se construyen sistemas automáticos de reconocimiento de locutor independientes para cada unidad lingüística, permitiendo así un análisis detallado de la capacidad discriminativa de las frecuencias formantes en cada unidad. Las frecuencias formantes en unidades lingüísticas son ampliamente utilizadas en la comparación de voz forense, pero hasta la fecha no había estudios publicados en aplicaciones de gran escala como el que se presenta en este trabajo compendiado, donde se usan bases de datos estándar *de facto* en el ámbito del reconocimiento automático de locutor y se presentan resultados para un gran número de unidades lingüísticas. Además, se presenta un análisis comparativo de las

unidades más discriminativas por locutor, con el fin de observar si existen unidades particulares que permiten distinguir con suficiente precisión a la mayoría de los locutores, o si por el contrario cada locutor presenta particularidades que hacen que pueda distinguirse mejor de otros a partir de unidades específicas.

Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors. **Javier Franco-Pedroso** and Joaquin Gonzalez-Rodriguez. Proceedings of Odyssey 2016: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 2016 (to appear)

En esta publicación se parte de dos de los trabajos anteriores, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition" y "Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data", para profundizar en el desarrollo de técnicas de reconocimiento de locutor que conecten las aproximaciones forense y automática. En ella también se hace uso de i-vectors obtenidos a partir de frecuencias formantes en unidades lingüísticas, pero en este caso se usan como características de entrada a un modelo generativo que permite proporcionar relaciones de verosimilitud directamente calibradas. De esta forma, se evita la etapa de calibración final utilizada habitualmente en los sistemas automáticos, que suele requerir de datos adicionales (un recurso escaso en entornos forenses).

1.3.2. Otras publicaciones relacionadas

- *Segmentación automática de audio*

ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzín 2010 Evaluation. **Javier Franco-Pedroso**, Ignacio Lopez-Moreno, Doroteo T. Toledano and Joaquin Gonzalez-Rodriguez. Proceedings of FALA: "VI Jornadas en Tecnología del Habla" and "II Iberian SLTech Workshop", November 10-12th, 2010, Vigo (Spain), pp: 415-418

Esta publicación describe los sistemas desarrollados para las evaluaciones Albayzín 2010 de segmentación de audio (Butko and Nadeu [2011]) y de *diarización* de locutores (Zelenák and Hernando [2012]), organizadas por la Red Temática en Tecnologías del Habla. Para la segmentación automática de audio se planteó un sistema de segmentación basado en modelos por decodificación del flujo de audio mediante un HMM hergódico de 5 estados, en el que cada estado consistía en un GMM. Cada uno de estos GMMs constituía una de las clases acústicas a reconocer y fue entrenado de forma discriminativa respecto al resto de clases con el objetivo de maximizar la separación entre ellas, al ser mutuamente excluyentes. El sistema de *diarización*, por su parte, implementaba un sistema de agrupamiento jerárquico a partir de i-vectors extraídos

en ventanas de corta duración, seguido de una decodificación de Viterbi para alinear los cambios de locutor de forma más precisa.

- *Análisis perceptual en reconocimiento de locutor*

Calibration and weight of the evidence by human listeners: the ATVS-UAM submission to NIST Human-Aided Speaker Recognition 2010. Daniel Ramos, **Javier Franco-Pedroso** and Joaquin Gonzalez-Rodriguez. Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 22-27th, 2011, Prague (Czech Republic), pp: 5908-5911. doi: 10.1109/ICASSP.2011.5947706

De forma adicional a las habituales evaluaciones de reconocimiento de locutor mediante sistemas automáticos organizadas por el NIST, en el año 2010 se planteó también una evaluación de reconocimiento de locutor que permitía la interacción de personas con los sistemas automáticos (National Institute of Standards and Technology [2010]), o incluso que las tareas de reconocimiento fueran llevadas a cabo únicamente por personas (en un conjunto de datos reducido). El grupo de investigación ATVS tomó parte en esta modalidad repartiendo las pruebas de reconocimiento entre un conjunto de personas no expertas que podían escuchar las grabaciones y visualizar su forma de onda y espectrograma a la hora de puntuar el parecido entre los locutores. Esta publicación recoge los resultados obtenidos y su comparación con los proporcionados por sistemas automáticos en el mismo conjunto de datos, así como un análisis de los efectos de la calibración en las puntuaciones proporcionadas por las personas, exponiendo así el problema de la evaluación de evidencias llevado a cabo por personas de forma perceptual.

What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. Joaquín Gonzalez-Rodriguez, Juana Gil, Rubén Pérez and **Javier Franco-Pedroso**. Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop, June 16-19th, 2014, Joensuu (Finland), pp: 33-40

En este trabajo se presenta un análisis perceptual llevado a cabo por fonetistas sobre parejas de locuciones erróneamente identificadas como pertenecientes al mismo locutor por un sistema de reconocimiento automático. El objetivo del análisis es determinar si existen elementos perceptuales claramente discordantes entre dichas parejas de locuciones que son ignorados por las técnicas del estado del arte en reconocimiento automático, con el fin de desarrollar características cuantitativas o detectores automáticos a partir de dichos elementos que puedan combinarse con los sistemas clásicos de reconocimiento de locutor.

- *Caracterización de locutores en unidades lingüísticas*

Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition. **Javier Franco-Pedroso**, Fernando Espinoza-Cuadros and Joaquin Gonzalez-Rodriguez. Proceedings of IberSPEECH 2012: "VII Jornadas en Tecnología del Habla" and "III Iberian SLTech Workshop", November 21-23th, 2012, Madrid (Spain), pp: 20-29. doi: 10.1007/978-3-642-35292-8_3

Este trabajo constituye una ampliación de la publicación compendiada "Fine-grained automatic speaker recognition using cepstral trajectories in phone units" en la que se construyen sistemas automáticos de reconocimiento de locutor a partir de unidades lingüísticas de mayor duración como los difonemas. Así mismo, se analizan nuevas estrategias de selección y combinación de unidades lingüísticas, cuya fusión da lugar a un rendimiento mejor que el del sistema de referencia.

Formant Trajectories in Linguistic Units for Text-Independent Speaker Recognition. **Javier Franco-Pedroso**, Fernando Espinoza-Cuadros and Joaquin Gonzalez-Rodriguez. Proceedings of the 2013 International Conference on Biometrics (ICB), June 4-7th, 2013, Madrid (Spain), pp: 1-6. doi: 10.1109/ICB.2013.6613001

Esta publicación representa uno de los trabajos intermedios entre las publicaciones compendiadas "Fine-grained automatic speaker recognition using cepstral trajectories in phone units" y "Linguistically-constrained formant-based i-vectors for automatic speaker recognition", en la que se caracteriza a los locutores a partir las trayectorias temporales de las frecuencias formantes en unidades lingüísticas, y se estudia su aplicabilidad al ámbito forense.

- *Otros trabajos en reconocimiento de locutor*

Speaker Clustering for Variability Subspace Estimation. Alicia Lozano-Díez, Ivan Gomez-Piris, **Javier Franco-Pedroso**, Javier Gonzalez-Dominguez and Joaquin Gonzalez-Rodriguez. Proceedings of IberSPEECH 2014: "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop", November 19-21th, 2014, LasPalmas de Gran Canaria (Spain), pp: 61-70

Los sistemas de reconocimiento de locutor en el estado del arte hacen uso de técnicas que modelan la variabilidad intra-locutor e inter-locutor. Para ello, es necesario disponer de bases de datos etiquetadas que permitan estimar ambos tipos de variabilidad. Sin embargo, puede darse el caso de que no se disponga de etiquetas de locutor para la base de datos de desarrollo; un ejemplo de este caso sería el de una aplicación para reconocimiento de locutores a partir de grabaciones de Youtube donde, a lo sumo, podrían separarse las grabaciones en segmentos donde sólo aparezca un único locutor gracias a técnicas de *diarización*. Posteriormente, debería determinarse cuáles de estos segmentos pertenecen al mismo locutor mediante técnicas de *clusternig*. Este trabajo

presenta un análisis de técnicas de *clusternig* con este objetivo, cuyos resultados se aplicaron en la evaluación de reconocimiento de locutor del NIST de 2014 (National Institute of Standards and Technology [2013]).

Bloque 2

Resultados, discusión y conclusiones

2.1. Resultados y discusión

En esta Sección se resumen y discuten los principales resultados obtenidos en los trabajos troncales de la Tesis. Mientras los detalles de implementación y resultados específicos (en términos de valores de rendimiento) se dejan para las publicaciones compendiadas, aquí se tratan los resultados desde una perspectiva global, destacando su importancia en relación con los objetivos planteados.

Estos resultados se han agrupado en tres grandes bloques atendiendo a los principales objetivos planteados en la Tesis:

- en primer lugar, se resumen los resultados obtenidos en segmentación de audio por medio de detectores calibrados, y se discute su aplicabilidad en el ámbito forense.
- en segundo lugar, se resumen y discuten los resultados obtenidos tras la revisión y propuesta de nuevas técnicas de cálculo de LRs directamente a partir de características (*feature-based* LRs).
- finalmente, se resumen los estudios realizados en reconocimiento automático de locutor e idioma mediante distintas aproximaciones a la caracterización fonético-acústica, y se discuten los resultados obtenidos en su aplicación al ámbito forense mediante distintas aproximaciones (*feature-based* y *score-based* LRs).

2.1.1. Segmentación de audio mediante combinación de detectores calibrados

Con motivo de la participación del Grupo de Reconocimiento Biométrico ATVS en la evaluación de segmentación de audio Albayzín 2014 (Ortega *et al.* [2014]), organizada por la Red Temática de las Tecnologías del Habla (RTTH), se diseñó e implementó un sistema de segmentación de audio basado en detectores de clases acústicas calibrados, cuya descripción se incluye en la publicación compendiada Franco-Pedroso *et al.* [2014] (Sección 3.1). En esta evaluación, el objetivo era detectar los segmentos de audio en que podían aparecer cualquiera de las siguientes

clases acústicas, solapadas o no: voz, ruido y música. Los resultados de este sistema en la citada evaluación se incluyen en la publicación compendiada Castán *et al.* [2015] (Sección 3.2).

La aproximación habitual de los sistemas de segmentación de audio es dividir el flujo de audio completo en segmentos disjuntos que presenten un contenido homogéneo (voz aislada, voz sobre música, etc.) mediante un proceso de clasificación en clases mutuamente excluyentes. De esta manera, el sistema debe diseñarse desde el inicio considerando todas las posibles combinaciones en que pueden presentarse las clases acústicas consideradas; en el caso de esta evaluación: voz, voz+música, voz+ruido, voz+música+ruido, música, música+ruido, ruido y silencio. Para ello, debe entrenarse un modelo por cada combinación o un clasificador que considere todas estas sub-clases mutuamente excluyentes, lo que hace los sistemas difícilmente escalables si en un momento dado quiere considerarse alguna clase acústica adicional o simplemente detectar de forma separada casos especiales de alguna de las anteriores (por ejemplo, distintos idiomas).

El sistema desarrollado, en cambio, se basa en sistemas independientes cuyo objetivo es detectar cada una de las clases acústicas propuestas por separado: un detector de voz, uno de música y uno de ruido. Así, cada uno se encarga de detectar la presencia de la clase en cuestión, independientemente de que también se presente alguna de las otras, generando una segmentación o etiquetado paralelo como muestra la Figura 2.1.

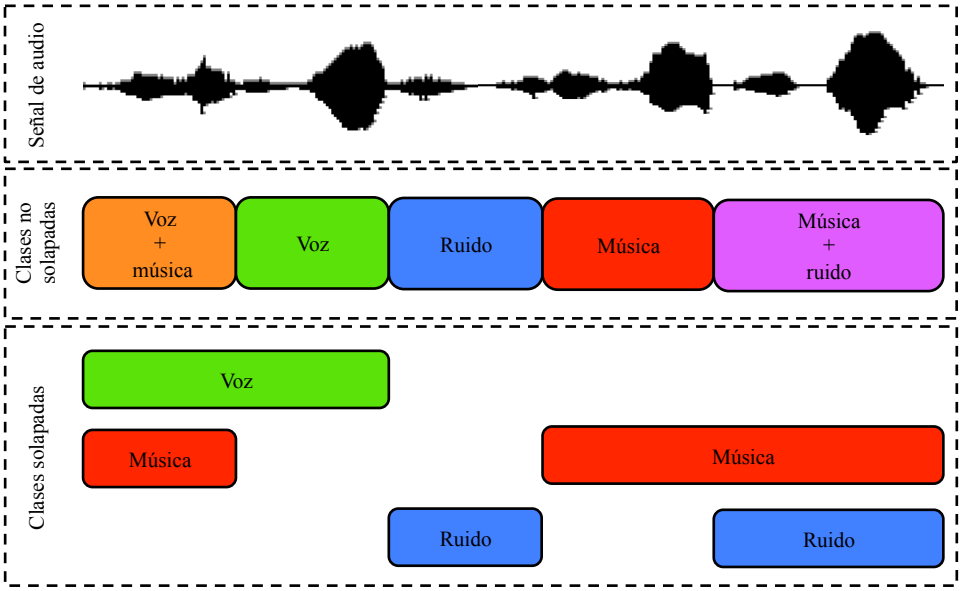


Figura 2.1: *Ejemplo de etiquetado de una señal de audio (arriba) en clases no solapadas (en medio) y en clases solapadas (abajo).*

Esta aproximación constituye un sistema fácilmente escalable, pues cada detector puede mejorarse o sustituirse por varios más especializados sin afectar al resto. Por ejemplo, el detector de voz puede sustituirse por dos detectores, uno de voz de hombre y otro de voz de mujer, sin necesidad de volver a entrenar el resto de detectores; en este caso, los detectores de ruido y música seguirán detectando ruido y música en presencia (o no) de voz, con independencia de que sea de hombre o mujer.

Por otra parte, la salida de cada detector en forma de LR permite la aplicación de este tipo de sistemas en el ámbito forense de forma que, por ejemplo, pueda tenerse en cuenta en el proceso de evaluación de evidencias la influencia de la presencia ruido o música en las muestras a evaluar. En este caso, las hipótesis asociadas a la relación de verosimilitud son:

- en la trama de audio se presenta la clase acústica c (H_c).
- en la trama de audio no se presenta la clase acústica c ($\overline{H_c}$).

donde $c = \{v, m, r\}$ representa a las clases acústicas antes mencionadas (**v**oz, **m**úsica y **r**uido). Es decir, sea x_t una trama de audio en un instante dado t, cada uno de los detectores proporciona la siguiente relación de verosimilitud:

$$LR_t^c = \frac{p(x_t|H_c)}{p(x_t|\overline{H_c})} \tag{2.1}$$

Al incorporar el proceso de calibración, el resultado de cada detector puede además combinarse fácilmente con el resto gracias al marco probabilístico que suponen los LRs, lo que permite transformar la segmentación en clases solapadas a una segmentación en clases no solapadas. Así, en la publicación compendiada Castán *et al.* [2015] (Sección 3.2) se incluyen, además de los resultados para cada sistema individual, una serie de experimentos de fusión entre los sistemas de los cuatro laboratorios participantes en la evaluación. Para ello, fue necesario proporcionar puntuaciones por trama para cada una de las sub-clases mutuamente excluyentes: voz, voz+música, voz+ruido, etc. Aunque el sistema de segmentación desarrollado no consideraba sub-clases no solapadas, las puntuaciones para estos casos fueron fácilmente derivadas a partir de los LRs obtenidos por cada detector. Por ejemplo, en el caso de la sub-clase no solapada voz+música, los LRs se obtuvieron mediante:

$$LR_t^{vm} = \frac{p(x_t|H_v \cap H_m)}{p(x_t|\overline{H_v} \cap \overline{H_m})} = \frac{p(x_t|H_v) \cdot p(x_t|H_m)}{p(x_t|\overline{H_v}) \cdot p(x_t|\overline{H_m})} = LR_t^v \cdot LR_t^m \tag{2.2}$$

pues en el esquema de segmentación propuesto la presencia de una clase acústica es independiente de que se presente cualquiera de las otras. Aunque el rendimiento del sistema diseñado por el autor no estuvo entre los sistemas ganadores, los experimentos de fusión demuestran que la aproximación seguida proporciona información muy complementaria a aquellos, pues la fusión que obtenía mejores resultados era una de las que incluía el sistema diseñado por el autor.

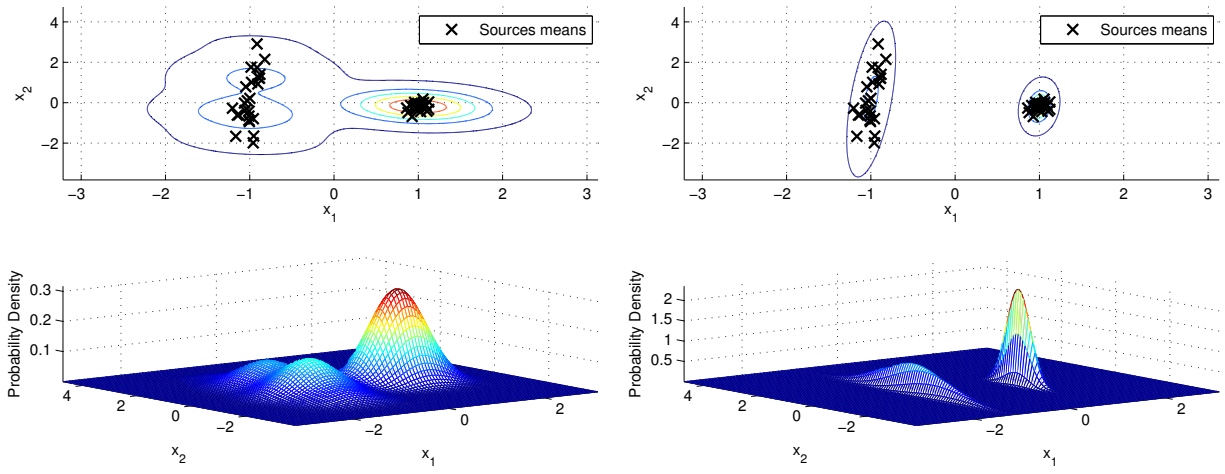
2.1.2. Cálculo de relaciones de verosimilitud a partir de datos multivariados

Una de las características de las aplicaciones forenses es la escasez de datos adicionales a los del caso a evaluar, que presenten condiciones similares y sean de relevancia para el caso en cuestión (lo que se denomina *población de referencia*). Esto dificulta su separación en conjuntos independientes para entrenar, por una parte, un sistema automático que permita la obtención de puntuaciones de similitud (*scores*), y por otra, un proceso de calibración posterior para su transformación a LRs (*score-based* LRs). Por ello, en la publicación compendiada Franco-Pedroso

et al. [2016] (Sección 3.3) se aborda el problema de la obtención de LRs en entornos forenses directamente a partir de las características medidas sobre las muestras a evaluar (*feature-based* LRs).

Una técnica habitualmente utilizada para este fin es la presentada en Aitken and Lucy [2004], donde se aplica un modelo generativo probabilístico que modela la variabilidad que se presenta, por una parte, entre muestras correspondientes a la misma fuente (variabilidad intra-fuente), y por otra, la que aparece entre muestras de distintas fuentes (variabilidad inter-fuente). Para la primera de ellas, se asume una distribución gaussiana común a las distintas fuentes. Para la segunda, se evalúa la distribución de las medias correspondientes a los conjuntos de muestras de cada fuente; en función de esta distribución, se aportan soluciones para el caso gaussiano y, cuando no se cumple esta condición, se aproxima mediante una función de densidad *kernel*. Estas variabilidades se estiman sobre la denominada *población de referencia* (o *background population*).

En la publicación compendiada Franco-Pedroso *et al.* [2016] (Sección 3.3), se revisa esta técnica y se propone para el segundo caso (distribución inter-fuente no gaussiana) el uso de modelos de mezcla de gaussianas (*gaussian mixture model*, GMM) para aproximar la distribución inter-fuente. Los análisis realizados muestran que la aproximación *kernel* tiende a sobrestimar la densidad de probabilidad en una amplia zona del espacio cuando las distintas fuentes se agrupan en zonas separadas. Los GMMs, en cambio, se ajustan mejor a los datos de esta naturaleza, como puede verse en la Figura 2.2.



(a) KDF

(b) GMM (2 componentes)

Figura 2.2: *Funciones densidad de probabilidad KDF y GMM para la variabilidad inter-objeto en un conjunto de datos sintético.*

El resultado de este mejor ajuste a la distribución de los datos en la población de referencia es una mejor calibración de los LRs obtenidos, medida por medio del *log-likelihood ratio cost* o C_{llr} (van Leeuwen and Brümmer [2007]). Así mismo, por medio de las curvas APE (van Leeuwen and Brümmer [2007]), se observa que la mejora se produce para distintos valores de las

probabilidades a priori. Además, esta mejora es consistente entre las diferentes bases de datos forenses empleadas (tintas, fragmentos de cristal y pinturas de coches).

Aunque no se destaca en la publicación compendiada anteriormente citada, otra de las ventajas de la aproximación mediante GMMs es el menor coste computacional que conlleva. La aproximación KDF puede verse como una suma equiponderada de funciones gaussianas, cada una centrada en la media de las muestras de una fuente; para los conjuntos de datos utilizados en el citado estudio, esto implica evaluar decenas de funciones gaussianas. La aproximación GMM, en cambio, al agrupar varias fuentes por componente, sólo necesita evaluar un número de funciones gaussianas igual al de componentes utilizadas; en el caso de los conjuntos de datos utilizados, del orden de unidades. La Tabla 2.1 muestra una comparativa entre ambos métodos para los conjuntos de datos utilizados en la publicación compendiada Franco-Pedroso *et al.* [2016] (Sección 3.3).

Base de datos	**KDF**	**GMM**
Tintas (40 fuentes)	40	1
Fragmentos de cristal (62 fuentes)	62	4
Pinturas de coche (36 fuentes)	36	5

Tabla 2.1: *Número de funciones gaussianas a evaluar para las aproximaciones KDF y GMM.*

Finalmente, es de especial importancia la incorporación de este tipo de técnicas al reconocimiento automático de locutor para su aplicación en entornos forenses. En la publicación compendiada Franco-Pedroso and Gonzalez-Rodriguez [2016a] (Sección 3.7), se utiliza este modelo probabilístico (asumiendo distribución gaussiana de la variabilidad inter-fuente) en un sistema de reconocimiento de locutor que incorpora caracterización acústico-fonética. Las muestras de entrada al modelo son, en este caso, i-vectors (Dehak *et al.* [2011]) obtenidos a partir de las repeticiones de una unidad lingüística determinada en una grabación del locutor a modelar. De esta forma, el sistema proporciona directamente LRs calibrados, evitando así el habitual proceso de calibración que requeriría de datos adicionales para su entrenamiento.

2.1.3. Caracterización acústico-fonética de locutor e idioma

2.1.3.1. Reconocimiento de idioma

Aunque la principal motivación de la caracterización acústico-fonética son las aplicaciones forenses del reconocimiento automático de locutor, en esta Tesis también se ha explorado su uso para el reconocimiento automático de idioma por dos motivos principales:

- por una parte, existen gran cantidad de ejemplos en la literatura científica que muestran que la información de alto nivel es muy complementaria a los sistemas en el estado del arte basados en características acústicas de bajo nivel.
- por otra, la identificación de idioma mediante sistemas que incorporen información fonética

puede proporcionar información útil en el estudio de las lenguas, por ejemplo para la caracterización de dialectos o establecer relaciones entre distintos idiomas.

En la publicación compendiada Gonzalez-Dominguez *et al.* [2010a] (Sección 3.4) se describe el sistema automático de reconocimiento de idioma desarrollado por el Grupo de Reconocimiento Biométrico ATVS para la evaluación de reconocimiento de idioma organizada por el NIST en el año 2009 (National Institute of Standards and Technology [2009]). Dicho sistema se basa en la combinación de varios sub-sistemas, uno de los cuales (indicado como ATVS4 en la citada publicación) es, a su vez, la combinación de varios sistemas *fonotácticos*.

Un sistema *fonotáctico* modela, a partir de la transcripción proporcionada por un reconocedor fonético, la frecuencia de aparición de fonemas o determinadas secuencias de ellos (llamadas n-gramas), caracterizando así a un idioma particular en base a los sonidos más habituales en esa lengua. Esta combinación (reconocedor fonético más modelo estadístico del idioma) se conoce como *Phone Recognizer Language Modelling* (PRLM). La Figura 2.3 muestra un ejemplo de verificación de idioma con un sistema de estas características, mediante un esquema de comparación respecto a un modelo universal (*universal background model*, UBM). La aproximación usada en el trabajo compendiado, en cambio, utiliza directamente las frecuencias de repetición como características de entrada a un clasificador SVM, lo que se conoce como *Phone-SVM*.

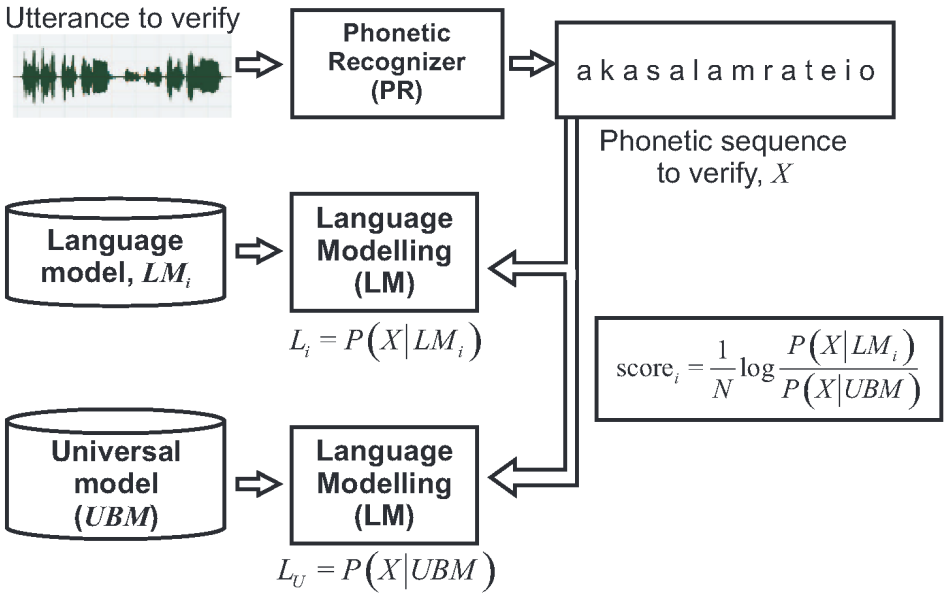


Figura 2.3: *Proceso de verificación mediante un esquema PRLM.*

La particularidad de este tipo de sistemas es que la codificación fonética usada no tiene porqué coincidir con la del idioma a reconocer; por ejemplo, puede usarse un reconocedor fonético de inglés para crear un modelo de español (u otros idiomas) en base a esa codificación fonética. En este contexto, el reconocedor fonético se usa para definir los distintos sonidos a considerar en el modelado del idioma, por lo que pueden usarse reconocedores fonéticos de distintos idiomas.

De hecho, el procedimiento más habitual es usar varios de estos sistemas en paralelo, cada uno basado en un reconocedor fonético (idioma) distinto, lo que se conoce como *Parallel PRLM*. Así, distintos reconocedores fonéticos aportan distintas caracterizaciones acústico-fonéticas, lo que enriquece el sistema global.

En la publicación compendiada Gonzalez-Dominguez *et al.* [2010a] (Sección 3.4) se muestra cómo un sistema formado por la combinación de 10 sub-sistemas *fonotácticos* obtenía resultados de identificación equivalentes, e incluso mejores en alguna de las distintas condiciones de evaluación, a los de un sistema acústico en el estado del arte. Además, combinando los dos anteriores, se lograba mejorar el rendimiento frente a cualquiera de ellos por separado, demostrando que ambas aproximaciones son complementarias.

2.1.3.2. Reconocimiento de locutor

Los principales trabajos realizados por el autor en el ámbito del reconocimiento automático de locutor (Secciones 3.5, 3.6 y 3.7) tienen en común el modelado de éste de forma independiente para distintas unidades lingüísticas. De esta forma, se adapta el proceso de reconocimiento automático a uno de los procedimientos habitualmente usados en entornos forenses, en los que las comparaciones de voz se realizan entre unidades lingüísticas equivalentes. Para ello, se hace uso de sistemas de reconocimiento automático de voz, con el objetivo de delimitar los intervalos de tiempo en los que se producen las distintas realizaciones de las unidades lingüísticas de interés. A partir de estas transcripciones, se filtra el flujo de características extraídas de forma que el modelo de locutor se construya sólo a partir de las asociadas a la unidad de interés, como muestra la Figura 2.4.

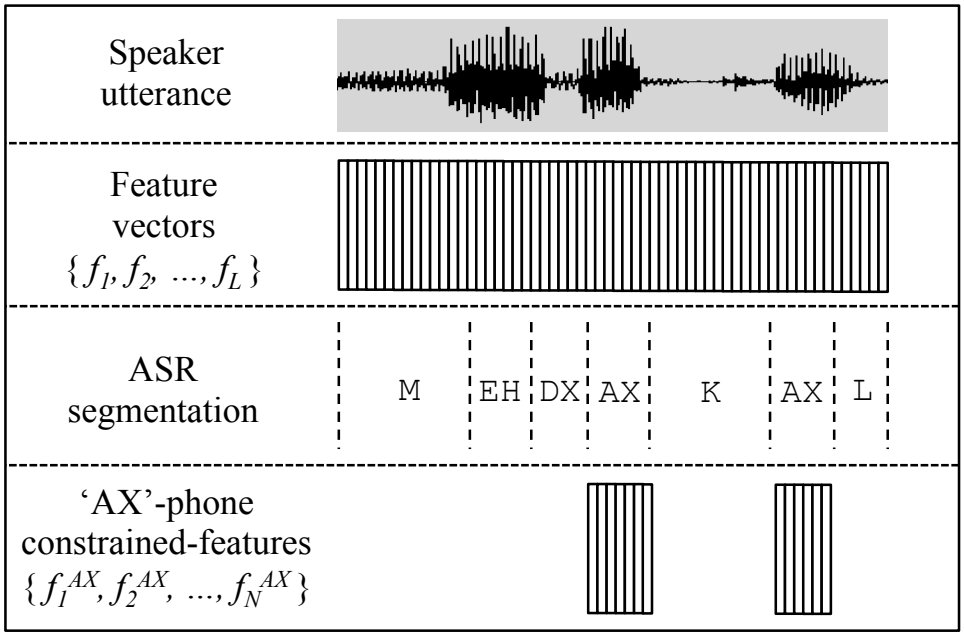


Figura 2.4: *Filtrado de características mediante transcripciones automáticas.*

Gracias a los trabajos previos del Grupo de Reconocimiento Biométrico ATVS en el ámbito de las aplicaciones forenses del reconocimiento de locutor, los autores fueron invitados a escribir un capítulo de libro para un compendio de trabajos forenses como homenaje al fonetista Phil Rose. La publicación compendiada Franco-Pedroso *et al.* [2012c] (Sección 3.5) constituye así uno de los primeros trabajos en que se aborda la aplicabilidad a entornos forenses del reconocimiento automático de locutor en base a unidades lingüísticas.

En esta publicación, las características asociadas a las unidades lingüísticas consisten en la codificación de la trayectoria seguida por los coeficientes MFCC a lo largo del fonema, concatenando finalmente las trayectorias de las distintas dimensiones en un único vector. A partir de estas características, se construyen sistemas de reconocimiento GMM-UBM (Reynolds *et al.* [2000]) independientes para cada uno de los fonemas analizados, seguidos de una etapa de calibración, como muestra la Figura 2.5 para un fonema determinado. De esta forma, fue posible analizar las capacidades discriminativas y de calibración de distintos fonemas, proporcionando además la correspondencia entre la codificación fonética utilizada por el sistema automático y el alfabeto fonético internacional para facilitar la interpretación de resultados por parte de los expertos en fonética.

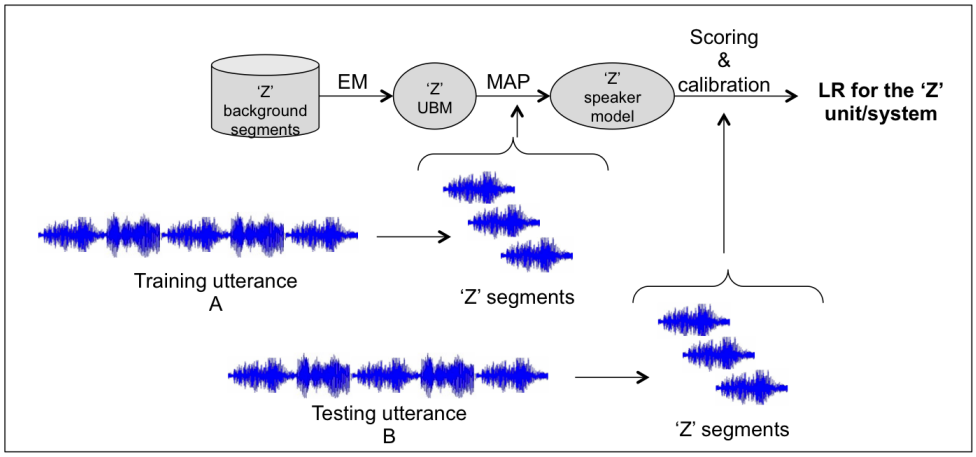


Figura 2.5: *Sistema automático de reconocimiento de locutor en base a las características del fonema 'Z'.*

Los resultados muestran que, en promedio, existen grandes diferencias de rendimiento dependiendo del fonema en base al cual se realice el reconocimiento de locutor, aunque las tasas de error en todos los casos están muy por encima de un sistema equivalente con modelado de locutor independiente de fonema (sistema de referencia). No obstante, se demuestra que es posible calibrar los resultados por fonema, obteniendo así LRs informativos utilizables en entornos forenses. Por otra parte, se observa que la combinación de varios de estos sistemas dependientes de fonema permite integrar satisfactoriamente la información discriminativa de locutor repartida entre ellos, llegando a rendimientos equivalentes a los del sistema de referencia.

A raíz de esta publicación, se profundizó en el estudio de este tipo de sistemas extendiendo el análisis a unidades lingüísticas más largas como los difonemas (Franco-Pedroso *et al.* [2012a]).

Aunque cabría esperar que el mayor recorrido de las trayectorias en difonemas podría proporcionar más poder de discriminación entre locutores , sucede que la frecuencia de aparición de difonemas es mucho menor que la de fonemas, lo que reduce la cantidad de datos disponible para el entrenamiento de modelos dependiente de unidad. No obstante, se observa que para algunos difonemas particulares, el rendimiento puede ser tan bueno como el de los mejores fonemas, lo que invita a pensar que ciertas transiciones entre fonemas incorporan información muy discriminativa. Así mismo, se profundizó en el análisis de estrategias de selección de unidades y técnicas de fusión para su combinación (2.6), lo que dio lugar a rendimientos superiores a los del sistema de referencia cuando se combinaban distintos fonemas y difonemas.

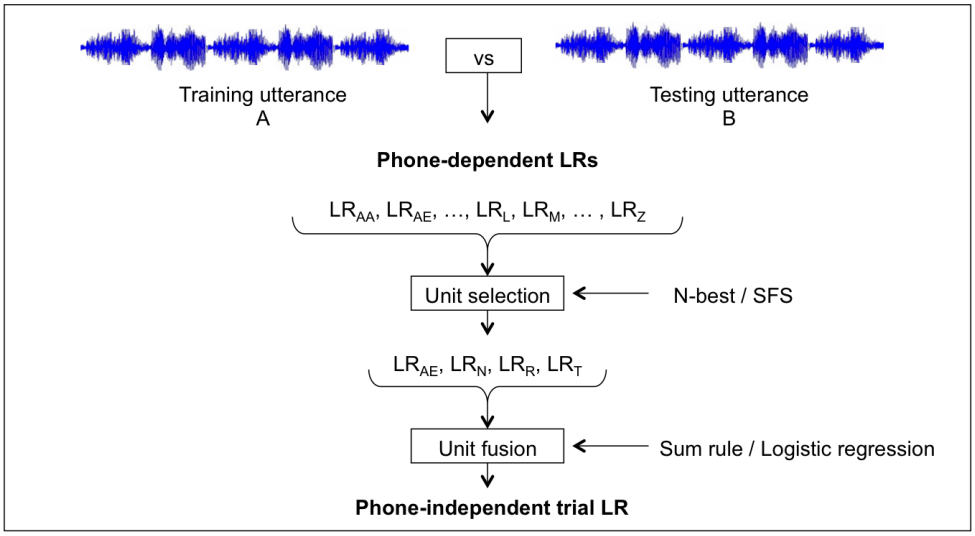


Figura 2.6: *Proceso de combinación de unidades en un único LR por comparación.*

En una publicación posterior (Franco-Pedroso *et al.* [2013]) se empleó el mismo esquema de verificación sobre trayectorias de frecuencias formantes calculadas de forma automática, incorporando así características fácilmente interpretables y ampliamente utilizadas en entornos forenses. En este caso, aunque el rendimiento por unidad era similar al de trayectorias de MFCCs para los fonemas, los difonemas presentaban, en promedio, mayores tasas de error. Así mismo, el rendimiento obtenido por la mejor combinación de unidades era inferior al del sistema de referencia GMM-UBM. Sin embargo, la combinación de ambas aproximaciones daba lugar a tasas de error más bajas que cuando se combinaba el sistema de referencia con los basados en trayectorias de MFCCs, demostrando que ambos tipos de características son muy complementarios a pesar de estar ambos basados información frecuencial.

Aunque las trayectorias en unidades lingüísticas demostraron aportar información muy útil en la caracterización acústico-fonética del locutor, el esquema de codificación y concatenación en un único vector por repetición de una unidad lingüística presenta dos inconvenientes importantes a la hora de aplicar un modelado estadístico. Por una parte, reduce la cantidad de vectores disponibles para el entrenamiento, y por otra, aumenta la dimensión de estos, agravando el problema anterior. Esto dificultaba la aplicación de técnicas de modelado en el estado del arte como

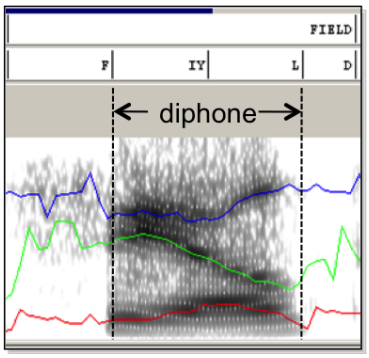


Figura 2.7: *Trayectorias de formantes en un difonema.*

los i-vectors (Dehak *et al.* [2011]), que requieren de gran cantidad de datos para la estimación de subespacios de variabilidad.

Por ello, en la publicación compendiada Franco-Pedroso and Gonzalez-Rodriguez [2016b] (Sección 3.6) se adopta un esquema diferente para incorporar la información temporal de las frecuencias formantes, aplicando los conocidos como coeficientes *delta*, ampliamente utilizados en procesado de voz para incorporar la información temporal en un entorno localizado de los coeficientes cepstrales. De esta forma, se consiguió aplicar de forma eficiente un esquema de verificación basado en i-vectors (Figura 2.8), cuyos resultados mejoraban significativamente a los obtenidos mediante trayectorias de formantes y un esquema de verificación GMM-UBM.

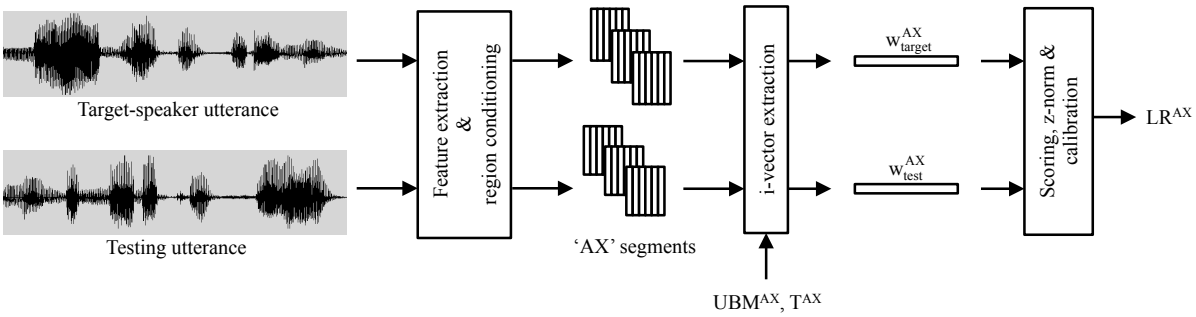


Figura 2.8: *Esquema de verificación mediante i-vectors basados en las frecuencias formantes del fonema 'AX'.*

Además del análisis de rendimiento promedio por unidad realizado en publicaciones anteriores, en este trabajo se analizó además el rendimiento de las distintas unidades para locutores específicos. Este análisis demuestra que, aunque existen determinadas unidades lingüísticas con mejor rendimiento promedio entre distintos locutores, cada locutor presenta ciertas particularidades que hacen que las unidades que presentan mejor rendimiento para él sean distintas de las de otros locutores. Es decir, para cada locutor existe un conjunto determinado de unidades lingüísticas en base a las cuales es más fácil distinguirlo de otros. Además se demuestra que, si el reconocimiento de esos locutores se basara únicamente en la unidad lingüística con mejor rendimiento para ellos, se obtendrían tasas de error tan bajas como las del sistema de referencia

(basado también en i-vectors, pero a partir de características cepstrales y sin modelado por unidad).

Finalmente, cabe destacar de nuevo la publicación compendiada Franco-Pedroso and Gonzalez-Rodriguez [2016a] (Sección 3.7). En ella, usando el mismo esquema de extracción de i-vectors por unidad que en la publicación anterior, se aplican técnicas de cálculo de LRs directamente a partir de dichos i-vectors. Aunque los LRs así obtenidos (*feature-based* LRs) presentan pérdidas de calibración ligeramente mayores que en el caso de aplicar un paso específico de calibración (*score-based* LRs), estas son suficientemente bajas, con la ventaja de no necesitar datos adicionales para el entrenamiento de este proceso. Además, el rendimiento por unidad mejora en promedio respecto a la aproximación previa.

2.2. Resumen de contribuciones específicas

A continuación se listan las contribuciones específicas de esta Tesis, agrupadas por temática (algunas publicaciones son comunes a distintas temáticas):

- **Revisiones bibliográficas:**
  1. Técnicas de extracción de características mediante sistemas automáticos de reconocimiento de voz: Franco-Pedroso and Gonzalez-Rodriguez [2016b]
  2. Uso de frecuencias formantes para reconocimiento locutor: Franco-Pedroso and Gonzalez-Rodriguez [2016b]
  3. Estrategias de segmentación automática de audio: Castán *et al.* [2015]
  4. Técnicas de cálculo de LRs a partir de datos multivariados: Franco-Pedroso *et al.* [2016]
  5. Reconocimiento automático de idioma: Gonzalez-Dominguez *et al.* [2010a]
- **Nuevos métodos:**
  1. Incorporación de información dinámica de frecuencias formantes en unidades lingüísticas mediante coeficientes *delta*: Franco-Pedroso and Gonzalez-Rodriguez [2016b],
  2. Modelado de variabilidad inter-fuente mediante GMMs para el cálculo de LRs a partir de datos multivariados: Franco-Pedroso *et al.* [2016]
  3. Cálculo de LRs a partir de características en reconocimiento automático de locutor: Franco-Pedroso and Gonzalez-Rodriguez [2016a]
  4. Segmentación de audio mediante combinación de detectores calibrados: Franco-Pedroso *et al.* [2014]
- **Aplicación a otras disciplinas forenses:**
  1. Cálculo de LRs a partir de datos multivariados en fragmentos de cristal, tintas y pinturas de coche: Franco-Pedroso *et al.* [2016].

- **Nuevas técnicas de caracterización del locutor:**

  1. Técnicas basadas en trayectorias de coeficientes cepstrales en unidades lingüísticas: Franco-Pedroso *et al.* [2012c], Franco-Pedroso *et al.* [2012a]
  2. Técnicas basadas en trayectorias de frecuencias formantes en unidades lingüísticas: Franco-Pedroso *et al.* [2013]
  3. Técnicas basadas en valores instantáneos de frecuencias formantes en unidades lingüísticas: Franco-Pedroso and Gonzalez-Rodriguez [2016b], Franco-Pedroso and Gonzalez-Rodriguez [2016a]

- **Nuevos estudios experimentales:**

  1. Análisis de la capacidad discriminativa y propiedades de calibración en unidades lingüísticas: Franco-Pedroso *et al.* [2012c], Franco-Pedroso *et al.* [2012a], Franco-Pedroso *et al.* [2013], Franco-Pedroso and Gonzalez-Rodriguez [2016b], Franco-Pedroso and Gonzalez-Rodriguez [2016a]
  2. Reconocimiento de locutor mediante análisis perceptual: Gonzalez-Rodriguez *et al.* [2014], Ramos *et al.* [2011]
  3. Técnicas de selección y combinación de sistemas: Gonzalez-Dominguez *et al.* [2010a], Franco-Pedroso *et al.* [2012c], Franco-Pedroso *et al.* [2012a], Franco-Pedroso *et al.* [2013], Franco-Pedroso and Gonzalez-Rodriguez [2016b], Franco-Pedroso and Gonzalez-Rodriguez [2016a]

- **Difusión de resultados:**

  1. Foros de reconocimiento automático:
  2. Reconocimiento automático de idioma: Gonzalez-Dominguez *et al.* [2009], Franco-Pedroso *et al.* [2010], Franco-Pedroso *et al.* [2012a], Franco-Pedroso *et al.* [2013], Franco-Pedroso *et al.* [2014], Gonzalez-Rodriguez *et al.* [2014], Franco-Pedroso and Gonzalez-Rodriguez [2016a]
  3. Foros de ciencias forenses: Franco-Pedroso *et al.* [2012b], Franco-Pedroso [2013]

- **Mejoras de la capacidad discriminativa en sistemas automáticos:**

  1. Contribuciones a la mejora de los sistemas automáticos de reconocimiento de locutor e idioma del Grupo de Reconocimiento Biométrico ATVS: Gonzalez-Dominguez *et al.* [2010a], Gonzalez-Dominguez *et al.* [2010b], Khoury *et al.* [2013] Lozano-Diez *et al.* [2014]
  2. Contribuciones a la mejora de los sistemas automáticos de segmentación de audio y *diarización* de locutores: Franco-Pedroso *et al.* [2010], Franco-Pedroso *et al.* [2014]

2.3. Conclusiones

Resumiendo, los principales resultados y contribuciones obtenidos en esta Tesis son:

- en el ámbito del reconocimiento de idioma, se ha demostrado que las aproximaciones basadas en sistemas *fonotácticos* son tan eficaces en esta tarea como las basadas en información acústica de bajo nivel, y muy complementarias a ellas.
- en el ámbito del reconocimiento de locutor, se han estudiado y validado varias aproximaciones a la caracterización acústico-fonética, siendo de especial interés las basadas en frecuencias formantes por el carácter interpretable de estas características. Los análisis realizados en base a este tipo de modelado han servido para demostrar que cada locutor tiene particularidades que hacen que sea más fácil distinguirlos de otros en base a determinadas unidades. Así mismo, se ha analizado la aplicabilidad de este tipo de sistemas en entornos forenses.
- se han estudiado y desarrollado técnicas de cálculo de relaciones de verosimilitud directamente a partir de las características, y se ha validado su uso tanto en reconocimiento de locutor a partir de caracterizaciones acústico-fonéticas como en bases de datos de otras disciplinas forenses.
- se han diseñado sistemas de segmentación automática integrables en el ámbito forense gracias a una aproximación basada en detectores calibrados. Además, esta aproximación proporciona una gran flexibilidad que permite distintos tipos de etiquetado (clases solapadas o no-solapadas) que pueden usarse indistintamente dependiendo de la aplicación final, y supone un sistema fácilmente escalable al no considerar clases mútuamente excluyentes.
- finalmente, se ha contribuido a la divulgación de aproximaciones que combinan técnicas tradicionales y automáticas tanto en foros de acústica-fonética como en los propios del reconocimiento automático.

Referencias

C. G. G. Aitken and D. Lucy. Evaluation of trace evidence in the form of multivariate data. 53(1):109–122, Feb. 2004. ISSN 0035-9254 (print), 1467-9876 (electronic). 6, 16

N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. In *Computer Speech and Language*, volume 20, pages 230 – 275, 2006. 2

T. Butko and C. Nadeu. Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011(1):1–10, 2011. ISSN 1687-4722. URL `http://dx.doi.org/10.1186/1687-4722-2011-1`. 8

D. Castán, D. Tavarez, P. Lopez-Otero, J. Franco-Pedroso, H. Delgado, E. Navas, L. Docio-Fernández, D. Ramos, J. Serrano, A. Ortega, and E. Lleida. Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–9, 2015. URL `http://dx.doi.org/10.1186/s13636-015-0076-3`. 14, 15, 23

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011. ISSN 1558-7916. 2, 17, 22

European Network of Forensic Science Institutes. ENFSI Guideline for Evaluative Reporting in Forensic Science, June 2015. URL `https://www.unil.ch/esc/files/live/sites/esc/files/Fichiers%202015/ENFSI%20Guideline%20Evaluative%20Reporting`. 2

J. Franco-Pedroso. Formant trajectories in linguistic units for text-independent speaker recognition. In *ENFSI FSAAWG (European Network of Forensic Science Institutes – Forensic Speech and Audio Analysis Working Group) 15th Meeting, Helsinki (Finland)*, September 2013. 24

J. Franco-Pedroso, F. Espinoza-Cuadros, and J. Gonzalez-Rodriguez. *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, chapter Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition, pages 20–29. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012a. ISBN 978-3-642-35292-8. URL `http://dx.doi.org/10.1007/978-3-642-35292-8_3`. 20, 24

J. Franco-Pedroso, F. Espinoza-Cuadros, and J. Gonzalez-Rodriguez. Formant trajectories in linguistic units for text-independent speaker recognition. In *Proceedings of the International Conference on Biometrics (ICB 2013), 4-7 June, 2013, Madrid, Spain*, pages 1–6, 2013. 21, 24

J. Franco-Pedroso and J. Gonzalez-Rodriguez. Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors. To appear in IEEE Odyssey 2016: The Speaker and Language Recognition Workshop, 2016, June 2016a. 17, 23, 24

J. Franco-Pedroso and J. Gonzalez-Rodriguez. Linguistically-constrained formant-based i-vectors for automatic speaker recognition. *Speech Communication*, 76:61 – 81, February 2016b. ISSN 0167-6393. URL `http://www.sciencedirect.com/science/article/pii/S0167639315001302`. 22, 23, 24

J. Franco-Pedroso, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and D. Ramos. Fine-grained automatic speaker recognition using cepstral trajectories in phone units. In *The International Association for Forensic Phonetics and Acoustics (IAFPA) 2012 Annual Conference, Santander (Spain)*, pages 415–418, August 2012b. 24

J. Franco-Pedroso, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and D. Ramos. *Quantitative approaches to problems in linguistics. Studies in honour of Phil Rose*, chapter Fine-grained automatic speaker recognition using cepstral-trajectories in phone units Fine-grained automatic Speaker recognition using cepstral-trajectories in phone units Fine-grained automatic Speaker recognition using cepstral-trajectories in phone units, pages 185–196. LINCOM GmbH 2012, 2012c. 20, 24

J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez. Atvs-uam system description for the audio segmentation and speaker diarization albayzin 2010 evaluation. In *Proceedings of FALA 2010: "VI Jornadas en Tecnología del Habla" and "II Iberian SLTech Workshop"*, pages 415–417, November 2010. 24

J. Franco-Pedroso, D. Ramos, and J. Gonzalez-Rodriguez. Gaussian mixture models of between-source variation for likelihood ratio computation from multivariate data. *PLoS ONE*, 11(2):1–25, 02 2016. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0149958`. 15, 16, 17, 23

J. Franco-Pedroso, E. G. Rincon, D. Ramos, and J. Gonzalez-Rodriguez. Atvs-uam system description for the albayzin 2014 audio segmentation evaluation. In *IberSPEECH 2014: "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop"*, pages 247–252, Las Palmas de Gran Canaria (Spain), November 2014. 13, 23, 24

E. Gold and P. French. International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, 18(2), 2011. ISSN 1748-8893. URL `https://journals.equinoxpub.com/index.php/IJSLL/article/view/11992`. 1

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. Toledano, and J. Gonzalez-Rodriguez. Multilevel and session variability compensated language recognition: Atvs-uam systems at nist lre 2009. *Selected Topics in Signal Processing, IEEE Journal of*, 4(6):1084–1093, Dec 2010a. ISSN 1932-4553. 18, 19, 23, 24

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano, and J. Gonzalez-Rodriguez. Atvs-uam nist sre 2010 system. In *Proceedings of FALA 2010*, November 2010b. 24

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T.Toledano, and J. Gonzalez-Rodriguez. Multilevel and channel compensated language recognition: Atvs system at nist lre 2009. In *I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language*, September 2009. 24

J. Gonzalez-Rodriguez. Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014). *Loquens*, 1(1):1–15, January 2014. 1

J. Gonzalez-Rodriguez, J. Gil, R. Pérez, and J. Franco-Pedroso. What are we missing with i-vectors? a perceptual analysis of i-vector-based falsely accepted trials. In *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, pages 33–40, 2014. 1, 24

J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. 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*, 15(7):2104–2115, Sept 2007. ISSN 1558-7916. 2

E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkcnafet, L. Mazaira Fernandez, M. Diez, J. Kosmala, H. Khemiri, T. Cipr, R. Saeidi, M. Gunther, J. Zganec-Gros, R. Candil, F. Simoes, M. Bengherabi, A. Alvarez Marquina, M. Penagarikano, A. Abad, M. Boulayemen, P. Schwarz, D. Van Leeuwen, J. Gonzalez-Dominguez, M. Neto, E. Boutellaa, P. Gomez Vilda, A. Varona, D. Petrovska-Delacretaz, P. Matejka, J. Gonzalez-Rodriguez, T. Pereira, F. Harizi, L. Rodriguez-Fuentes, L. El Shafey, M. Angeloni, G. Bordel, G. Chollet, and S. Marcel. The 2013 speaker recognition evaluation in mobile environment. In *Biometrics (ICB), 2013 International Conference on*, pages 1–8, June 2013. 24

T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010. 2

A. Lozano-Diez, I. Gomez-Piris, J. Franco-Pedroso, J. Gonzalez-Dominguez, and J. Gonzalez-Rodriguez. Speaker clustering for variability subspace estimation. In *Proceedings of IberSPEECH 2014, Las Palmas de Gran Canaria, Spain*, November 2014. 24

G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, and C. Goemans Dorny. Interpol survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263:92–100, April 2016. URL `http://dx.doi.org/10.1016/j.forsciint.2016.03.044`. xxi, xxi, 1, 2, 3

National Institute of Standards and Technology. The 2009 NIST Language Recognition Evaluation Plan (LRE09), 2009. URL `http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf`. 6, 18

National Institute of Standards and Technology. The NIST Year 2010 Speaker Recognition Evaluation Plan, 2010. URL `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf`. 9

National Institute of Standards and Technology. The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge, 2013. URL `http://nist.gov/itl/iad/mig/upload/sre-ivectorchallenge_2013-11-18_r0.pdf`. 11

A. Ortega, D. Castan, A. Miguel, and E. Lleida. The Albayzin 2014 Audio Segmentation Evaluation, 2014. URL `http://iberspeech2014.ulpgc.es/images/segm_eval.pdf`. 5, 13

D. Ramos, J. Franco-Pedroso, and J. Gonzalez-Rodriguez. Calibration and weight of the evidence by human listeners. the atvs-uam submission to nist human-aided speaker recognition 2010. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5908–5911, May 2011. 24

D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia. Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1–8, June 2006. 2

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000. URL `http://dx.doi.org/10.1006/dspr.1999.0361`. 20

P. Rose. *Forensic Speaker Identification*. Forensic Science. Taylor and Francis, 2002. 2

D. van Leeuwen and N. Brümmer. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In C. Müller, editor, *Speaker Classification I*, volume 4343 of *Lecture Notes in Computer Science*, pages 330–353. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-74186-2. 16

G. Zadora, A. Martyna, D. Ramos, and C. Aitken. *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data.* Wiley, 2014. 6

H. Zelenák, Martin Schulz and J. Hernando. Speaker diarization of broadcast news in albayzin 2010 evaluation campaignand. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):1–9, 2012. URL `http://dx.doi.org/10.1186/1687-4722-2012-19`. 8

Bloque 3

Copia completa de los trabajos compendiados

3.1. ATVS-UAM System Description for the Albayzin 2014 Audio Segmentation Evaluation

Título: "ATVS-UAM System Description for the Albayzin 2014 Audio Segmentation Evaluation"
Autores: Javier Franco-Pedroso, Elena Gomez Rincon, Daniel Ramos and Joaquin Gonzalez-Rodriguez
Congreso: "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop" (IberSPEECH 2014), 19-21 de Noviembre, 2014, Las Palmas de Gran Canaria (España)
Páginas: 247-252

ATVS-UAM System Description for the Albayzin 2014 Audio Segmentation Evaluation

Javier Franco-Pedroso, Elena Gomez Rincon, Daniel Ramos and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Universidad Autonoma de Madrid (UAM). Spain
http://atvs.ii.uam.es
javier.franco@uam.es,elena.gomezr@estudiante.uam.es,daniel.ramos@uam.es,joaquin.gonzalez@uam.es

Abstract. This document describes the audio segmentation system developed by the ATVS – Biometric Recognition Group, at Universidad Autonoma de Madrid (UAM), for the Albayzin 2014 Audio Segmentation Evaluation (ASE). This system is based on three independent GMM-UBM acoustic-class detectors based on MFCC-SDC features. Each acoustic-class detector ('mu', 'no', 'sp') evaluates test recordings in a frame-by-frame manner, and the score-streams are filtered and calibrated previous to the detect-decision stage. Although the performance of the independent acoustic-class detectors is far from being perfect in terms of EER, the resulting audio segmentation systems achieves low miss (7.9%), false alarm (10.6%) and class error (3.0%) rates, given a final 21.43% SER on our development subset.

Keywords: audio segmentation, MFCC-SDC, GMM-UBM, calibration

1 Introduction

In contrast to our previous participation in Albayzin ASE campaigns (the 2010 edition [2]), this year we present a lighter but more robust system that avoids the overfitting introduced by Maximum Mutual Information discriminative training when the available data is scarce. Moreover, the system developed fits better the approach followed in this campaign by the organizers to the problem of evaluating automatic segmentation systems [3]: instead of labeling non-overlapping segments of (maybe overlapped) different acoustic classes, the presence of each acoustic class should be independently annotated in different segments (maybe overlapped with other acoustic classes). Although the problem can be solved from both perspectives (training different models for each possible acoustic-classes combination as we did in 2010 campaign), considering one independent detector for each acoustic class provides a more scalable solution and avoids the constraints regarding the available data for training the acoustic models.

The system developed consists in three independent acoustic-class detectors (speech –'sp'-, music -'mu'-, and noise –'no'-) based on the classical GMM-UBM

adfa, p. 1, 2011.

framework [4]. Each detector performs a frame-by-frame scoring of the test recordings, obtaining one log-likelihood stream per acoustic class. These score-streams are smoothed through a mean filter over a sliding window in order to deal with the high variability of frame-scores. Finally, smoothed frame-scores are independently calibrated by means of a linear logistic regression trained on a subset of the development dataset.

The remainder of this paper is organized as follows. Section 2 describes the feature extraction process. Sections 3 and 4 describe, respectively, the acoustic-class modeling and the acoustic-class detection stage. Section 5 explains the experimental protocol followed, and shows the results obtained in our development subset. Finally, Section 6 summarizes the key points of our submission, exposes the computational requirements and draws some conclusions.

2 Feature Extraction

Shifted Delta Coefficients (SDC) [5] have been widely used in Language Recognition due to the fact that they capture the time dependency structure of the language better than the speed or acceleration coefficients (also known as delta and delta-delta). Similarly, SDC features are expected to capture the different time dependency of the music over the speech or noise. In fact, experiments carried out over a subset of the development tracks revealed that GMM-UBM detectors build from MFCC-SDC features outperform those trained on MFCC plus delta coefficients.

For both development and evaluation tracks, one feature vector was extracted every 10 ms by means of a 20 ms Hamming sliding window (50% overlap). For each window, 7 MFCC features (including C0) were computed from 25 Mel-spaced magnitude filters over the whole available spectrum (0-8000 Hz). These features have been mean-normalized, RASTA filtered and Gaussianized through a 3-second window. Finally, their SDC were computed on a 7-1-3-7 (N-D-P-K) configuration and concatenated with them in a 56-coefficient feature vector.

3 Acoustic-Class Modeling

Acoustic classes have been modeled adopting the classical GMM-UBM framework [4] widely used for speaker recognition. First, a 1024-component UBM was trained by means of a 1-iteration k-means initialization followed by a 5-iteration EM stage. For this purpose, one half of the development dataset provided was used (tracks 01-10). Secondly, acoustic-class models were MAP-adapted [4] from this UBM through 1 single iteration and using a relevance factor r=16. Again, tracks 01-10 were used also for this step.

For each acoustic class, training data were extracted from segments belonging to the same acoustic-class as appeared in the provided development labels. This means that, for instance speech segments may contain not only isolated speech but also any of the other acoustic classes overlapped with it. As we are aiming to develop an acoustic-class detector, our assumption is that the acoustic-class models should collect

their own acoustic class in any possible condition it may appear. On the other hand, segments where each class can be found isolated are very scarce in the database provided, so robust acoustic-class models cannot be trained from such small amount of data, as we found out in our preliminary experiments.

4 Acoustic-Class Detection Stage

Acoustic-class detection stage is based on a frame-by-frame scoring of the test track against every acoustic-class model. Frame-by-frame log-likelihoods are highly variable over time, as it can be seen on Figure 1. For a segment with an isolated acoustic-class, it is expected that the mean log-likelihood will converge to a stable value as long as more frames are incorporated, as it has been shown for the speaker recognition task in [6]. For this reason, these score-streams were smoothed through a mean filter over a sliding window in order to have a more stable frame-score that approaches the "true" score of the acoustic class present in the surrounding frames. Figure 2 shows the result of applying this mean filtering stage for a 700-frame sliding window. The window length was independently optimized for each acoustic-class detector, looking for the length that provides the best detection performance in terms of EER. Results are shown in Figure 3 for our development subset (tracks 11-15).

Finally, the frame-by-frame log-likelihoods were calibrated by means of a linear logistic regression implemented in FoCal toolkit [1]. One different logistic regression is used for each acoustic-class detector, all of them trained on the same development subset used for the window length optimization (tracks 11-15).

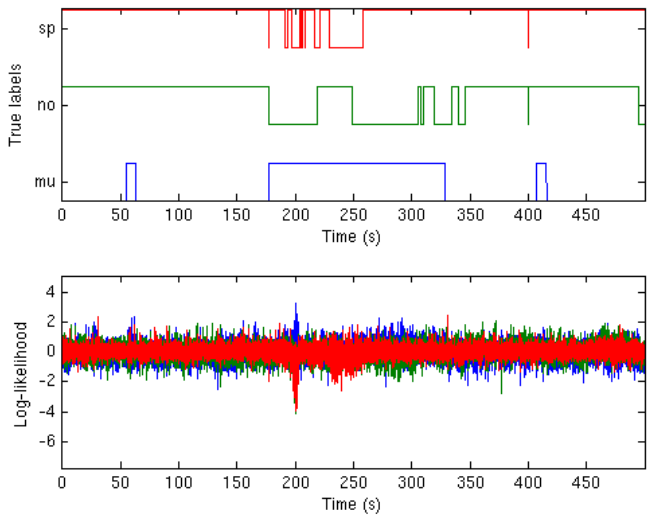


Fig. 1. Detail of the frame log-likelihoods for a 500-second segment of track11.

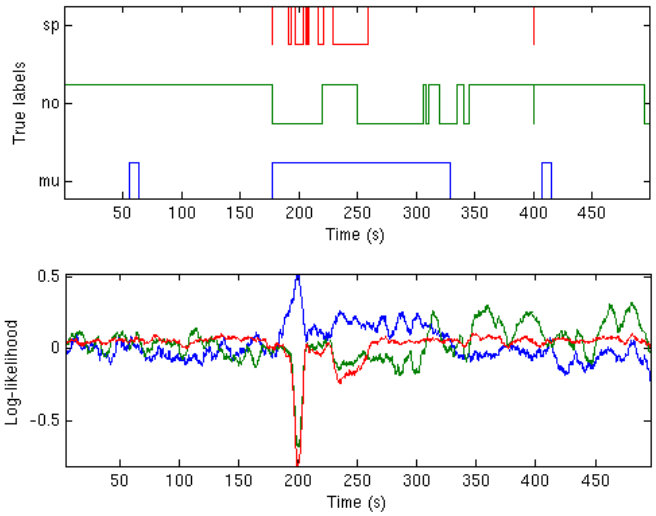


Fig. 2. Detail of the frame log-likelihoods for a 500-second segment of track11 after the mean filtering stage.

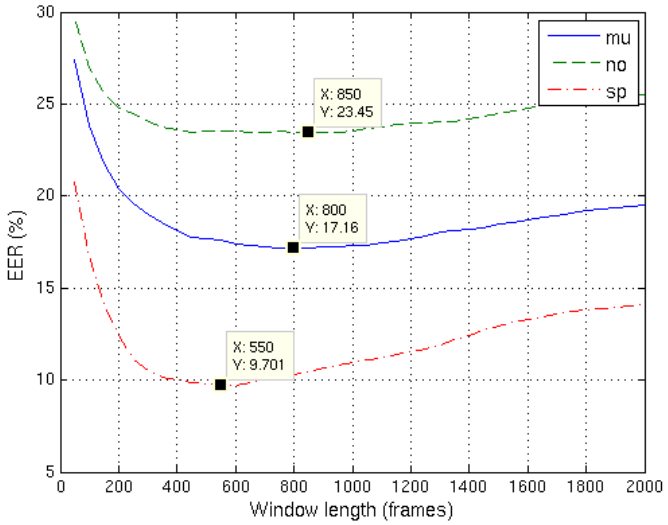


Fig. 3. EER as a function of the mean-filtering window-length, obtained for our development subset (tracks11-15). Best results are highlighted (X: window length, Y: EER).

5 Experimental setup and development results

Table 1 shows how the development data have been partitioned in order to be used for different purposes. One half of the development dataset has been devoted to train the acoustic models. From the remaining subset, one half has been used to find the optimum window length for the frame-scores mean-filtering, and the resulting frame-scores used to train the calibration rule; the final 5-track subset has been left apart in order to test the developed system.

Table 1. Dataset partitioning for system development.

Purpose	Track numbers
UBM training	01-10
Acoustic-class modeling	01-10
Window length optimization	11-15
Calibration training	11-15
Audio segmentation testing	16-20

Segmentation results obtained for our test subset (tracks 16-20) are shown in Table 2. As it can be seen, in spite of having acoustic-class detectors of relatively low detection performance (9.7% EER for 'sp', 17.2% EER for 'mu' and 23.4% EER for 'no'), the whole audio segmentation system achieves good performance compared with results shown in previous Albayzin ASE campaigns.

Table 2. Performance of the audio segmentation system: missed class time, false alarm class time, class error time and overall segmentation error, in seconds and percentages.

Error	Time (s)	% scored class time
Missed Class	2262.51	7.9
False Alarm Class	3057.21	10.6
Class error	853.85	3.0
Overall Segmentation Error	**21.43 %**	

6 Summary and conclusions

ATVS – Biometric Recognition Group has developed an efficient and light audio segmentation system. This system is based on three independent GMM-UBM acoustic-class detectors that can be developed and tuned independently. For instance, detectors in submitted systems make use of a different mean-filtering window-length and independent score-calibration rules, but they could be based in different features as well. Moreover, the adopted approach of modeling broad acoustic classes ('mu', 'no', 'sp') instead of the specific sub-classes given by all the possible combinations ('mu+no', 'sp+no', etc.) allows to develop a more robust system and avoids overfitting when the available training data is scarce. Finally, it can be seen in Table 3 that the computational requirements in terms of CPU time are very low, allowing the

testing to be run in 0.225xRT for each track. Experiments were carried out in a machine equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 16GB of RAM.

Table 3. Testing time per track (~60 min) for the different stages and total time as a real-time (xRT) factor.

Stage	Time
Feature extraction	19 secs
Frame-by-frame scoring	13 min
Scores filtering and calibration	5 sec
Total (xRT)	~0.225

Acknowledgement

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz, TEC2012-37585-C02-01).

References

1. Niko Brummer, FoCal: toolkit for evaluation, fusion and calibration of statistical pattern recognizers (2008). Online: http://sites.google.com/site/nikobrummer/focal
2. J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez, ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation. In Proceedings of FALA: VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop, 2010, pp. 415–418.
3. Alfonso Ortega, Diego Castan, Antonio Miguel, Eduardo Lleida. The Albayzin 2014 Audio Segmentation Evaluation. Online: http://iberspeech2014.ulpgc.es/images/segm_eval.pdf
4. Reynolds, D., Quatier, T., Dunn, R., Speaker Verification Using Adapted Gaussian Mixture Models Digital Signal Processing, vol. 10, 19–41 (2000).
5. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. Proc. ICSLP 2002, Sept. 2002, pp. 89-92.
6. Robbie Vogt and Sridha Sridharan, Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions. Lecture Notes in Computer Science Volume 5558, 2009, pp 454-463.

3.2. Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains

Título: "Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains"
Autores: Diego Castán, David Tavarez, Paula Lopez-Otero, Javier Franco-Pedroso, Héctor Delgado, Eva Navas, Laura Docio-Fernández, Daniel Ramos, Javier Serrano, Alfonso Ortega and Eduardo Lleida
Revista: EURASIP Journal on Audio, Speech, and Music Processing (Factor de impacto 5 años, 2014: 0.624, cuartil: Q4), Volumen 2015, Número 1, Diciembre de 2015
Páginas: 1-9
Editor: Springer
doi: 10.1186/s13636-015-0076-3

Castán *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:33
DOI 10.1186/s13636-015-0076-3

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains

Diego Castán[1*], David Tavarez[2], Paula Lopez-Otero[3], Javier Franco-Pedroso[4], Héctor Delgado[5], Eva Navas[2], Laura Docio-Fernández[3], Daniel Ramos[4], Javier Serrano[5], Alfonso Ortega[1] and Eduardo Lleida[1]

Abstract

Audio segmentation is important as a pre-processing task to improve the performance of many speech technology tasks and, therefore, it has an undoubted research interest. This paper describes the database, the metric, the systems and the results for the Albayzín-2014 audio segmentation campaign. In contrast to previous evaluations where the task was the segmentation of non-overlapping classes, Albayzín-2014 evaluation proposes the delimitation of the presence of speech, music and/or noise that can be found simultaneously. The database used in the evaluation was created by fusing different media and noises in order to increase the difficulty of the task. Seven segmentation systems from four different research groups were evaluated and combined. Their experimental results were analyzed and compared with the aim of providing a benchmark and showing up the promising directions in this field.

Keywords: Audio segmentation, Broadcast news, Albayzín-2014 evaluation

1 Introduction

Automatic audio segmentation aims at providing boundaries to delimit portions of audio with homogeneous acoustic content. The resulting segments are classified in different acoustic types according to the final application, such as different speakers, languages, speech/non-speech portions, or acoustic events among others. In most cases, automatic audio segmentation is considered a pre-processing tool to improve the performance of the subsequent system related with speech technologies. For example, in very large multimedia repositories, the speech is usually found along with music or environmental noise. The presence of these acoustic classes must be accurately labeled because it is critical for the subsequent systems to be successful. Thus, the development of accurate Audio Segmentation Systems is essential to allow post-processing systems, such as automatic speech recognition (ASR) or spoken document retrieval (SDR), to perform adequately in real-world environments.

Audio segmentation systems can address the problem in different fields or contexts. In the first works of automatic segmentation, the goal was the challenging segmentation of sports material and commercials. The studies focused on speech/music segmentation from radio stations as in [1] and [2] showing the importance of the audio segmentation to improve ASR systems. The following studies dealt with the recognition of broad classes to produce an adaptation of the ASR models. For example, Srinivasan [3] classified the audio of a video into mixed classes such as music with speech or speech with background noise using a combination of acoustic and perceptual features. Nowadays, most of the studies focus on the robust and generic segmentation of broad classes [4] and the segmentation of acoustic events [5] for audio retrieval in large multimedia databases.

A specific task with large multimedia databases is the segmentation of broadcast news (BN) recordings. This task is very challenging because the audio contains different kinds of sequences with a very heterogeneous style. Several international evaluation campaigns, such as the TREC NIST evaluations for SDR [6], the ESTER evaluation campaigns for rich transcription (RT) in French [7], and the COST278 evaluation for segmentation and

*Correspondence: dcastan@unizar.es
[1] ViVoLab, Universidad de Zaragoza, Zaragoza, Spain
Full list of author information is available at the end of the article

speaker clustering in a multi-lingual domain [8], have already been proposed to face this task in the past.

Nowadays, the amount of audio documents is exponentially increasing due to the audio-sharing websites or the audio-on-demand systems. Users around the world can upload and share their contents and, for that reason, the variability of the acoustic conditions is extremely high. As a result, systems must be able to adapt their role in high-variability data spaces, providing robust performance in different conditions. Due to the importance of audio segmentation and the need to develop robust systems capable of operating over a rich variety of audio conditions, the Albayzín-2014 campaign was proposed as an international evaluation to measure the performance of segmentation systems for different databases and different contexts. This segmentation evaluation, which is part of an open set of evaluations organized by the RTTH[1] every 2 years, compares systems and approaches from different research institutions in an independent way.

In contrast to previous evaluations such as Albayzín-2010 [9], where five unambiguous acoustic classes were defined, the Albayzín-2014 evaluation proposed the delimitation of the presence of speech, music and/or noise that can be found simultaneously. Another relevant difference was the composition of the database: while in previous evaluations the databases were composed of a unique BN media (TV in Albayzín-2010 mostly in Catalan [9] or radio in Albayzín-2012 [10] mostly in Spanish), the Albayzín-2014 database was a combination and fusion of three different databases with TV, radio, and noise recordings. This composition increased the difficulty of the task since the resulting database introduced more variability, presenting more realistic conditions over a wide variety of acoustic sources.

The remainder of the paper is organized as follows: the database and the metric used for Albayzín-2014 segmentation evaluation are presented in Section 2. Section 3 briefly describes the submitted systems. The results of the evaluation and the fusion of the systems are presented and discussed in Section 4. Finally, the summary and the conclusions are presented in Section 5.

2 Database and evaluation metric

The proposed evaluation consisted of segmenting a broadcast audio document and assigning labels for each segment indicating the presence of speech, music, and/or noise. That is, two or more classes could be found simultaneously in audio segments and the goal was to indicate if one, two, or the three aforementioned classes were present for a given time instant. For example, music could be overlapped with speech, or noise could be found in the background when someone was speaking. Therefore, the presence of these three classes involve the definition of eight non-overlapping classes: silence, speech, music, noise, speech with music, speech with noise, music with noise, and speech with music and noise. In this evaluation, Speech was present every time that a person was speaking but not in the background or singing. Music was understood in a general sense and noise was considered every time some acoustic content was present different than speech and music (including speech in the background, which usually comes from a crowd).

The goal was to segment and label audio documents indicating where speech, music, and/or noise were present. Unlike 2010 evaluation criteria [9], no prior classes were defined (*speech, music, speech with noise in background, speech with music in background, other*) and a multiple layer labeling approach was proposed instead. In summary, the goal was to segment the incoming audio into three (possibly overlapped) acoustic classes: speech, music, and noise, where the audio was drawn from different databases that have been merged or even overlapped, thus dramatically increasing the difficulty of the task with regard to previous evaluations.

2.1 Database

The database for this evaluation is a combination of three databases defined below:

The first dataset is the Catalan broadcast news database from the 3/24 TV channel proposed for the Albayzín-2010 Audio Segmentation Evaluation [9]. This database was recorded by the TALP Research Center of the Polytechnic University of Catalonia in 2009 under the Tecnoparla project [11] funded by the Government of Catalonia. The Corporació Catalana de Mitjans Audiovisuals (CCMA), owner of the multimedia content, allows its use for technology research and development. The database consists of around 87 h of recordings in which speech can be found 92 % of the time, music is present 20 % of the time, and noise in the background is present 40 % of the time. Another class called *others* was defined which can be found 3 % of the time. Regarding the overlapped classes, speech can be found along with noise 40 % of the time and along with music 15 % of the time.

The second dataset is the Aragón Radio database from the Corporación Aragonesa de Radio y Televisión (CARTV) which was used for the Albayzín-2012 Audio Segmentation Evaluation [10]. As the owner of the audio content, Aragón Radio and the Corporación Aragonesa de Radio y Televisión allow the use of these data for research purposes.

The last dataset is composed of environmental sounds from Freesound.org [12] and HuCorpus [13]. These sounds were merged with segments from the 3/24 TV and Aragón Radio databases.

All the data was supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency. The database includes approximately 35 h of audio: 20 h

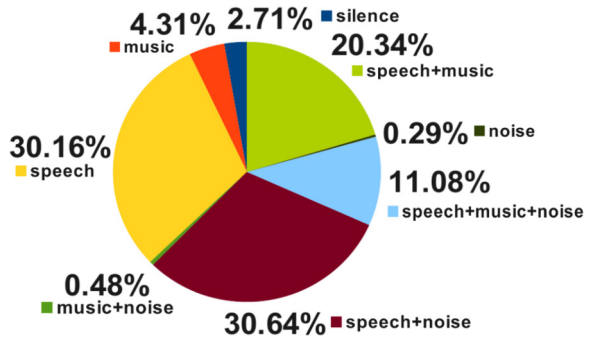


Fig. 1 Distribution of the different audio classes in the training set of the Albayzín-2014 evaluation

were used for the training set and 15 h were used for the test set. The distribution of the audio classes in the training set is presented in Fig. 1. The chart shows that the classes containing speech represent more than 92 % of the total time. There are two residual classes, isolated noise and music with noise, that represent less than 0.5 and 0.3 % of the total time, respectively.

2.2 Evaluation metric

As in the NIST RT Diarization evaluations [14], the segmentation error score (SER) was used to measure the performance of the proposed systems. SER is computed as the fraction of class time that is not correctly attributed to that specific class (speech, noise, or music). The SER score was computed over the entire file to be processed, including regions where more than one class was present (overlap regions).

The overall SER score is defined as the ratio of the overall segmentation error time to the sum of the durations of the segments that are assigned to each class in the file.

Given the dataset to evaluate Ω, each document is divided into contiguous segments at all "class change points" which occur each time any reference class (oracle) or system class (hypothesis) starts or ends. Thus, the set of active reference classes and/or system classes does not change during any segment. The segmentation error time for each segment n is defined as

$$E(n) = T(n)\left[\max\left(N_{\text{ref}}(n), N_{\text{sys}}(n)\right) - N_{\text{Correct}}(n)\right] \quad (1)$$

where $T(n)$ is the duration of segment n, $N_{\text{ref}}(n)$ is the number of reference classes that are present in segment n, $N_{\text{sys}}(n)$ is the number of system classes that are present in segment n, and $N_{\text{Correct}}(n)$ is the number of reference classes in segment n correctly assigned by the segmentation system.

$$\text{SER} = \frac{\sum_{n\in\Omega} E(n)}{\sum_{n\in\Omega} \left(T(n)N_{\text{ref}}(n)\right)} \quad (2)$$

The segmentation error time includes the amount of time that is assigned to the wrong class, missed class time, and false alarm class time:

- *Class error time*: The class error time is the amount of time that has been assigned to an incorrect class. This error can occur in segments where the number of system classes is greater than the number of reference classes but also in segments where the number of system classes is lower than the number of reference classes whenever the number of system classes and the number of reference classes are greater than zero.
- *Missed class time*: The missed class time refers to the amount of time that a class is present but not labeled by the segmentation system in segments where the number of system classes is lower than the number of reference classes.
- *False alarm class time*: The false alarm class time is the amount of time that a class has been labeled by the segmentation system but is not present in segments where the number of system classes is greater than the number of reference classes.

The forgiveness collar defines a no-score area around reference segment boundaries. Typically, the collar is 250 ms for speaker diarization tasks [14] and 1 s for segmentation tasks [9]. A forgiveness collar of 1 s, before and after each reference boundary, was considered in order to take into account both inconsistent human annotations and the uncertainty about when a class begins or ends. This collar is enough for the purpose of this segmentation task where the goal is to identify the areas with speech and their background to allow the adaptation of models for other systems as ASR. The implementation of the collar was provided by scoring tool of the NIST RT Diarization evaluations [14].

3 Segmentation systems

3.1 General description of audio segmentation systems

The general scheme of an audio segmentation system can be divided into two basic steps: the feature extraction method and the segmentation/classification strategy. Lavner in [15] and more recently Theodorou in [16] provide good reviews of the features and the classification methods used in the literature.

The acoustic feature extraction is the first step in an audio segmentation system. The audio input is divided into overlapping windows and, for each window, a feature vector is extracted. The feature vectors are descriptors used to distinguish the differences among classes in the time and frequency domains. Features can be grouped into two classes according to the time span they represent: *frame-based* and *segment-based*. *Frame-based* features are extracted within short periods of time (between

10 and 30 ms) and are commonly used in speech-related tasks where the signal can be considered stationary over that frame. *Mel Frequency Cepstrum Coefficients* (MFCC) or *Perceptual Linear Prediction* (PLP) coefficients are generally used as *frame-based* features as proposed in [17–22] among a great collection of works. *Frame-based* features have also been proposed for segmenting and classifying BN audio into broad classes. As an example, two pitch-density-based features are proposed in [23], the authors use *short-time energy* (STE) in [1, 24, 25], and *harmonic features* are used in [26–28]. The *frame-based* features can be directly used in the classifier. However, some classes are better described by the statistics computed over longer periods of time (from 0.5 to 5 s long). These characteristics are referred in the literature as *segment-based* features [29, 30]. For example, in [31], a content-based speech discrimination algorithm is designed to exploit the long-term information inherent in the modulation spectrum; and in [32], authors propose two segment-based features: the *variance of the spectrum flux* (VSF) and the *variance of the zero crossing rate* (VZCR).

Once the feature vectors are computed, the next step deals with the detection and the classification of the segments. The segmentation/classification strategies can be divided into two different groups depending on how the segmentation is performed. The first group detects the break-points in a first step and then classifies each delimited segment in a second step. We refer to them as *segmentation-and-classification* approaches but they are also known in the literature as *distance-based techniques*. These algorithms have the advantage that they do not need labels to delimit the segments because the segmentation is based on a distance metric estimated for adjacent segments. When the distance between two adjacent segments is greater than a certain threshold, a break-point is set and identified as an acoustic change-point. The resulting segments are clustered or classified in a second stage. The *Bayesian Information Criterion* (BIC) is a well-known distance-based algorithm. It is widely employed in many studies, such as [33], to generate a break-point for every speaker or environment/channel condition change in the BN domain and also, in [34] and [35], to identify mixed-language speech and speaker changes, respectively. The second group of segmentation/classification strategies is known as *segmentation-by-classification* or *model-based segmentation*. In contrast to the *segmentation-and-classification* algorithms, these algorithms classify consecutive fixed-length audio segments and, therefore, segment labels are required in a training step because each class of interest is described by a model. The segmentation is produced directly by the classifier as a sequence of decisions. This sequence is usually smoothed to improve the segmentation performance, since the classification of frames produces some spurious labels because adjacent frames are poorly considered.

A good and common approach to this procedure can be found in [36] where the author combines different features with a Gaussian Mixture Model (GMM) and a maximum entropy classifier. In [37], the authors use a factor analysis approach to adapt a universal GMM model to classify BN in five different classes. The final decisions of both systems are smoothed with a Hidden Markov Model (HMM) to avoid sudden changes.

Both segmentation/classification strategies were used by participants in the Albayzín-2014 Audio Segmentation Evaluation: three participating groups chose *segmentation-by-classification* algorithms with different model strategies and one participating group chose a *segmentation-and-classification* algorithm based on BIC for the first stage and on different classification systems for the second stage. A brief description of the features and the systems is given below.

3.2 Description of the participating systems

Four research groups participated in this evaluation with seven different systems: Aholab-EHU/UPV (University of the Basque Country), GTM-UVigo (University of Vigo), ATVS-UAM (Autonomous University of Madrid), and CAIAC-UAB (Autonomous University of Barcelona). Each participant had 3 months to design the segmentation system with the training data. After that time, participants were given 1 month to process the test data. The participants had to submit their results with hard-segmentation labels (in RTTM format from NIST) along with a technical description of the submitted systems. All participant teams had to submit at least a primary system but they could also submit up to two contrastive systems. Also, for fusion purposes, participants were required to submit the frame-level scores for each non-overlapping audio class. Groups are listed in the order in which their primary systems were ranked in the evaluation. A more detailed description of the systems can be found in the *Advances in Speech and Language Technologies for Iberian Languages* proceedings [38].

3.2.1 Group 1

This group presented a single primary system where two different *segmentation-by-classification* strategies were fused to build a robust system.

The first strategy consisted of a hidden Markov model (HMM) scheme with eight separate HMM models for each non-overlapping class: silence, speech, music, noise, speech with music, speech with noise, music with noise, and speech with music and noise. Thirteen MFCCs with first and second derivatives were used for the classification and each HMM had 3 states with 512 Gaussian components per state.

The second strategy consisted of a GMM presegmentation and a speech label refinement by means of i-vector classification via multilayer perceptron (MLP). Six GMMs with 32 components for silence, music, noise, clean speech, speech with noise, and speech with music were used in a Viterbi segmentation. Twelve MFCCs with first- and second-order derivatives were used for the classification (the energy-related coefficient was not used in this case). Once the speech segments were identified, the i-vector extraction process was carried out. A sliding window was used to extract the i-vectors corresponding to each speech segment. Then, an MLP was used to classify each i-vector as clean speech, speech with noise, speech with music, or speech with music and with noise.

The outputs of both subsystems were post-processed to discard too short segments. Finally, a label fusion algorithm based on the confusion matrices of the systems involved in the fusion was applied to combine the results of both subsystems and maximize the precision of the final labels.

3.2.2 Group 2

Group 2 presented a primary system and two contrastive systems, all of them with a *segmentation-and-classification* strategy.

The segmentation stage was common for all the systems and consisted of a Bayesian Information Criterion (BIC) approach using 12 MFCCs plus energy and featuring a false alarm rejection strategy: the occurrence of acoustic change-points was supposed to follow a Poisson process, and a change-point was discarded with a probability that varied in function of the expected number of occurrences in the time interval going from the previous change-point to the candidate change-point.

The classification stage was different for each system. The primary system was developed using i-vector representations of the segments obtained from the previous step with logistic regression classification. Perceptual linear prediction (PLP) analysis was used to extract 13 cepstral coefficients, which were combined with two pitch features and augmented with their delta features.

The classification in contrastive system 1 consisted of a Gaussian mean supervector representation of the segments obtained from the previous step through the adaptation of a Universal Background Model (UBM) with 256 components. Classification was performed employing a support vector machine (SVM) with a linear kernel. The feature vectors used in this classifier were 12 MFCCs plus energy as in the segmentation stage, augmented with their delta and delta-delta coefficients.

The contrastive system 2 used a classic GMM maximum likelihood classification with 512 components performed by doing MAP adaptation of a UBM with full-covariance matrices. The set of features was the same that was used in the primary system.

3.2.3 Group 3

Group 3 presented a single primary system based on three independent GMM-UBM detectors of broad acoustic classes (speech, music, and noise in every possible context) with a *segmentation-by-classification* strategy.

The system was based on MFCC feature vectors including shifted delta coefficients to capture the time dependency structure of the audio. Acoustic classes were modeled through 1024-component MAP-adapted GMMs. Each detector performed a frame-by-frame scoring obtaining one log likelihood stream per acoustic class. These score-streams were smoothed through an average filter over a sliding window in order to deal with the high variability of frame scores. Finally, the smoothed frame-level scores were independently calibrated for each acoustic class by means of linear logistic regression.

3.2.4 Group 4

Group 4 presented a primary system and a contrastive system with a *segmentation-by-classification* strategy for both of them.

The proposed system was based on a "binary key" (BK) modeling approach originally designed for speaker recognition [39] and later applied successfully in a speech activity detection task [40]. The approach provided a compact representation of a class model through a binary vector (vector only containing zeros and ones) by transforming the continuous acoustic space into a discrete binary one. This transformation was done by means of a UBM-like model called Binary Key Background Model (KBM). Once the binary representation of the input audio was obtained, subsequent operations were performed in the binary domain, and calculations mainly involve bit-wise operations between pairs of binary keys. Segment assignment was done by comparing each segment BK with the N BKs (previously estimated using the KBM and training data) for each of the N target audio classes. Two alternatives to compute the similarity between two binary keys were proposed, one for the primary system and other for the contrastive system, respectively.

4 Experimental results

This section presents and analyzes the results of the Albayzín-2014 Audio Segmentation Evaluation for all the primary and contrastive systems of each group.

Table 1 shows the segmentation error rate (as defined in Eq. 2) for the seven submitted systems. No system was trained with additional material apart from the audio provided for the evaluation. As can be seen from the table, both first (20.68 %) and second (20.80 %) best

Table 1 Segmentation error rate of participating systems

	Primary	Contrastive 1	Contrastive 2
Group 1	20.68	-	-
Group 2	20.80	29.13	22.52
Group 3	30.67	-	-
Group 4	31.59	33.93	-

systems obtained very similar performance even though the systems represent very different strategies to perform the segmentation: the primary system of group 1 is based on a fusion of two *segmentation-by-classification* approaches while the primary system of group 2 is based on a *segmentation-and-classification* approach. The primary systems of group 3 and group 4 also show similar performance (around 31 %), but they are far from the primary systems of groups 1 and 2. It is apparent that, for all groups, the primary systems outperform the contrastive systems, indicating that the choice of the main strategy of each group was done properly.

Figure 2 compares the misses (blue columns) and the false alarms (orange columns) of the participating systems for the overlapped acoustic classes individually (music, noise, and speech). Each system is represented by a *gX_Y* notation where *X* indicates the group index and *Y* indicates if the system is primary (*p*) or contrastive (*c*). The main source of the segmentation error comes from the noise detection, but the music detection also presents a considerably high error rate. This is because the music and noise classes are rarely presented alone but instead mixed with speech. Also, the lack of data for these isolated classes makes very difficult to train suitable models to detect them. Note that the two best systems (the primary systems of group 1 and group 2) have almost the same error rate coming from the detection of the noise class and both are much lower than those of the rest of the systems. The main difference between the systems submitted by groups 1 and 2 is that the former detects the music class better than the latter while the latter detects the speech classes slightly better than the former.

To accurately analyze the source of the errors, Fig. 3 presents the confusion matrices of the primary systems. The matrices show the percentage of the reference classes (rows) associated to hypothesized non-overlapping acoustic classes (columns). The classes are represented as SI for "silence," MU for "music," NO for "noise," SP for "speech," MN for "music+noise," SM for "speech+music," SN for "speech+noise," and SA for "speech+music+noise." The matrices clearly show that the most common errors are the confusions between "speech+music+noise" with "speech+noise" or "speech+music" and also between "speech+noise" and "speech." In addition, there is a common error in all the systems with "music+noise" being classified as "music." Note that the systems of group 2 and group 4 incur in a non-negligible error rate coming from the detection of the "silence" class since these systems do not implement a silence detector and, therefore, false alarms are produced.

A fusion of different systems usually improves the final result because the information comes from various sources [41]. For that purpose, the participants provided frame-level scores for each non-overlapping audio class for the training and test datasets. Table 2 shows the segmentation error rate when the scores of the primary systems are combined. The fusion was done with different combinations of the primary systems: group 1 and group 2 in the first row of the table; groups 1, 2, and 3 in the second row; group 1, 2, and 4 in the third row; and a combination of all the systems in the fourth row of the table. We used a set of techniques to combine the scores. Firstly, one Gaussian distribution is estimated with class-dependent full covariance and mean with maximum likelihood on the training data for each class. This technique is known as Gaussian Back-End (GBE) and the results are shown in the first column of the table. We trained the fusion model with the scores computed over the training dataset, and we used the test dataset to compute the SER. To smooth the decisions, a Viterbi algorithm was chosen to determine the maximum likelihood transitions among classes. The segments are delimited by the transitions given by the Viterbi algorithm (second column of Table 2). On the

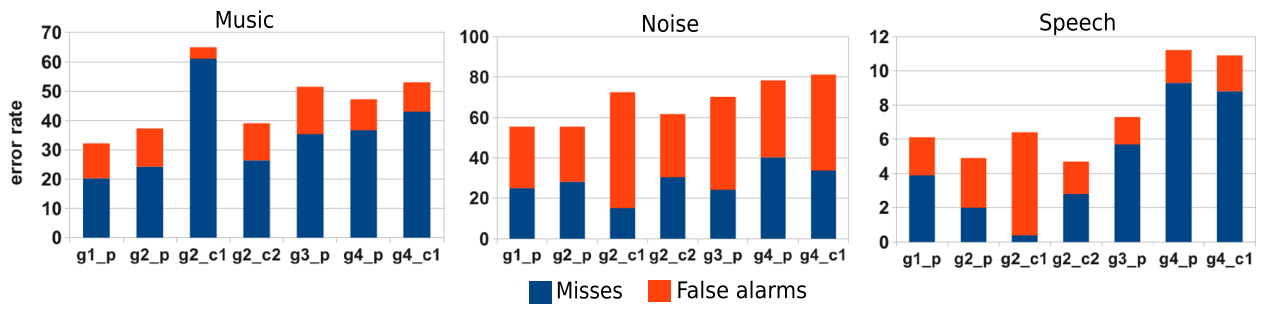


Fig. 2 Distribution of errors across the eight systems and for each acoustic class

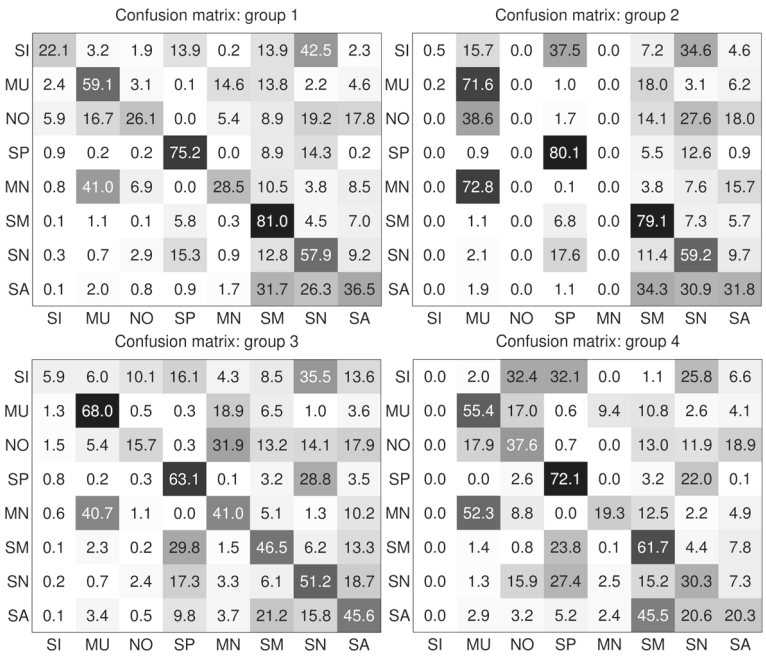


Fig. 3 Confusion matrices of the primary systems for the non-overlapped classes: SI (silence), MU (music), NO (Noise), SP (speech), MN (music and noise), SM (speech with music), SN (speech with noise), and SA (speech with music and noise)

other hand, we used the segments provided by the group 2 since the approach of this group is based on a previous unsupervised segmentation stage with BIC. In this case, we accumulated the log likelihood of each frame within the same segment, which was given the label of the class with the highest accumulated log likelihood. The results of this approach can be seen in the third column of the table. The best performance was attained when fusing the primary systems of groups 1, 2, and 3. Finally, the last column of the table gives us an idea about the performance of the fusion if the segmentation was perfect. It clearly shows a degradation between 4 and 5 % in the segmentation stage with regard to the oracle segmentation, because occasionally the delimitation boundaries among segments may be fuzzy. However, the fusion reduces the segmentation error rate for all the approaches compared with the winning primary system.

Table 2 Segmentation error rate of several score-level fusions of the primary systems. The result of the system G1 is 20.68 for comparison purposes

	GBE	GBE Viterbi	GBE AccumLLk GTM Seg.	GBE AccumLLk Oracle Seg.
G1+G2	19.60	19.41	19.30	14.36
G1+G2+G3	19.56	19.31	*19.16*	14.30
G1+G2+G4	19.94	19.77	19.64	15.58
G1+G2+G3+G4	19.86	19.67	19.62	15.31

5 Conclusions

This article presents the Albayzín-2014 Audio Segmentation Evaluation, including the main features of the database, an overview of the participating systems and evaluation and post-evaluation results. The new Albayzín-2014 audio segmentation database combines data from two different media (TV and radio), with added noises of diverse nature, thus increasing the difficulty of the task. Using this database an audio segmentation task was proposed, where the systems were required to identify the presence of speech, music and/or noise, either isolated or overlapped. The Albayzín-2014 Audio Segmentation Evaluation contributed to the evolution of the audio segmentation technology in broadcast news domains by providing a more general and realistic database, compared to those used in the Albayzín-2010 and -2012 Audio Segmentation Evaluations [10, 30]. The main features of the approaches and the results attained by seven segmentation systems from four different research groups have been presented and briefly analyzed. Three of the systems were based on a segmentation-and-classification strategy, while the rest of them were based on a segmentation-by-classification strategy.

Then, we presented seven segmentation systems and the results from four different research groups which participated in the Albayzín-2014 evaluation. The approaches and the results of each group were studied and compared. Three of the seven systems (from the same group) are based on a *segmentation-and-classification* strategy while

the rest of the systems are based on a *segmentation-by-classification* strategy. Most of the systems used common speech recognition features, such as MFCC, LFCC, or PLP.

The two best systems attained a segmentation error rate (SER) of around 20 %, following two different strategies but with a common classification approach based on i-vectors, showing the competitiveness of this technique. Both systems revealed that the main source of segmentation error was the detection of the noise class, mainly due to the low energy of noise signals. The results were analyzed using the non-overlapping classes through the confusion matrices of the primary systems. The matrices showed that the most common errors were the confusions between "speech+music+noise" with "speech+noise" or "speech+music" and also between "speech+noise" and "speech." Finally, the participating systems were combined under different approaches, yielding a relative improvement of up to 7.35 % SER.

Endnote

[1] Spanish Thematic Network on Speech Technologies: http://www.rthabla.es.

Abbreviations

AC: acoustic classes; BN: broadcast news; HMM: hidden Markov model; GMM: Gaussian mixture model.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work has been partially funded by the Spanish Government and the European Union (FEDER) under the project TIN2011-28169-C05-02 and supported by the European Regional Development Fund and the Spanish Government ('SpeechTech4All Project' TEC2012-38939-C03).

Author details

[1]ViVoLab, Universidad de Zaragoza, Zaragoza, Spain. [2]Aholab, Universidad del País Vasco, Bilbao, Spain. [3]Multimedia Technologies Group (GTM), AtlanTIC Research Center, Universidade de Vigo, Vigo, Spain. [4]ATVS, Universidad Autónoma de Madrid, Madrid, Spain. [5]CAIAC, Universitat Autònoma de Barcelona, Barcelona, Spain.

Received: 10 April 2015 Accepted: 16 November 2015
Published online: 01 December 2015

References

1. J Saunders, in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Real-time discrimination of broadcast speech/music, vol. 2 (IEEE Atlanta, 1996), pp. 993–996
2. E Scheirer, M Slaney, Construction and evaluation of a robust multifeature speech/music discriminator. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. **2**, 1331–1334 (1997)
3. S Srinivasan, D Petkovic, D Ponceleon, in *Proceedings of the Seventh ACM International Conference on Multimedia*. Towards robust features for classifying audio in the CueVideo system (ACM New York City, NY, 1999), pp. 393–400
4. S Kiranyaz, AF Qureshi, M Gabbouj, A generic audio classification and segmentation approach for multimedia indexing and retrieval. IEEE Trans. Audio Speech Lang. Process. **14**(3), 1062–1081 (2006)
5. Z Huang, Y-c Cheng, K Li, V Hautamaki, C-h Lee, in *Proc. Interspeech*. A blind segmentation approach to acoustic event detection based on i-vector (ISCA Lyon, 2013), pp. 2282–2286
6. NIST, TREC NIST Evaluations. http://www.itl.nist.gov/iad/mig//tests/sdr/ Accessed 23 Nov 2015
7. S Galliano, E Geoffrois, D Mostefa, in *Interspeech*. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news (ISCA Lisbon, 2005), pp. 3–6
8. J Zibert, F Mihelic, J Martens, H Meinedo, J Neto, L Docio, C Garcia-Mateo, P David, E Al, in *Interspeech*. The COST278 broadcast news segmentation and speaker clustering evaluation-overview, methodology, systems, results (ISCA Lisbon, 2005)
9. T Butko, CN Camprubí, H Schulz, in *II Iberian SLTech*. Albayzin-2010 audio segmentation evaluation: evaluation setup and results (FALA Vigo, 2010), pp. 305–308
10. A Ortega, D Castan, A Miguel, E Lleida, The Albayzin 2012 Audio Segmentation Evaluation (2012). http://dcastan.vivolab.es/wp-content/papercite-data/pdf/ortega2012.pdf Accessed 23 Nov 2015
11. Tecnoparla, Tecnoparla Project. http://www.talp.upc.edu/tecnoparla
12. F Font, G Roma, X Serra, in *Proceedings of the 21st ACM International Conference on Multimedia*. Freesound technical demo (ACM Barcelona, Spain, 2013)
13. G Hu, 100 non-speech environmental sounds. http://www.cse.ohio-state.edu/dwang/pnl/corpus/HuCorpus.html Accessed 23 Nov 2015
14. NIST, The 2009 (RT-09) Rich transcription meeting recognition evaluation plan. http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf Accessed 23 Nov 2015
15. Y Lavner, D Ruinskiy, A decision-tree-based algorithm for speech/music classification and segmentation. EURASIP J. Audio Speech Music Process. **2009**, 1–15 (2009)
16. T Theodorou, I Mporas, N Fakotakis, An overview of automatic audio segmentation. I.J. Inf. Technol. Comput. Sci. **1**, 1–9 (2014)
17. S Imai, in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Cepstral analysis synthesis on the mel frequency scale (IEEE Boston, 1983), pp. 93–96
18. R Vergin, D O'Shaughnessy, V Gupta, in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Compensated mel frequency cepstrum coefficients, vol. 1 (IEEE Atlanta, 1996), pp. 323–326
19. R Vergin, Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Trans. Speech and Audio Process. **7**(5), 525–532 (1999)
20. E Wong, S Sridharan, in *International Symposium on Intelligent Multimedia, Video and Speech Processing*. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification (Kowloon Shangri-La Hong Kong, 2001), pp. 95–98
21. M Hasan, M Jamil, M Rahman, in *International Conference on Computer and Electrical Engineering*. Speaker identification using Mel frequency cepstral coefficients (Dhaka, 2004), pp. 28–30
22. P Dhanalakshmi, S Palanivel, V Ramalingam, Classification of audio signals using AANN and GMM. Appl. Soft Comput. **11**(1), 716–723 (2011)
23. L Xie, W Fu, Z-H Feng, Y Luo, Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news. Multimedia Syst. **17**(2), 101–112 (2011)
24. D Li, I Sethi, N Dimitrova, T McGee, Classification of general audio data for content-based retrieval. Elsevier, Pattern Recogn. Lett. **22**, 533–544 (2001)
25. L Lu, H Zhang, H Jiang, Content analysis for audio classification and segmentation. IEEE Trans. Speech Audio Process. **10**(7), 504–516 (2002)
26. TL Nwe, H Li, in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Broadcast news segmentation by audio type analysis, vol. 2 (IEEE Philadelphia, 2005), p. 1065
27. A Hauptmann, R Baron, M Chen, in *Proc. TRECVID*. Informedia at TRECVID 2003: analyzing and searching broadcast news video (NIST Gaithersburg, 2003)
28. S Dharanipragada, M Franz, Story segmentation and topic detection in the broadcast news domain. DARPA Broadcast News Workshop, 1–4 (1999). http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa99/html/abstract.htm Accessed 23 Nov 2015
29. A Gallardo-Antolín, J Montero, Histogram equalization-based features for speech, music, and song discrimination. Signal Process. Lett. **17**(7), 659–662 (2010)
30. T Butko, C Nadeu, Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. EURASIP J. Audio Speech Music Process. **2011**(1), 1 (2011)

31. M Markaki, Y Stylianou, Discrimination of speech from nonspeech in broadcast news based on modulation frequency features. Speech Commun. **53**(5), 726–735 (2011)
32. R Huang, J Hansen, Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. IEEE Trans. Audio Speech Lang. Process. **14**(3), 907–919 (2006)
33. SS Chen, PS Gopalakrishnan, in *Proc. DARPA Broadcast News Workshop*. Speaker, environment and channel change detection and clustering via the Bayesian information criterion (Lansdowne, 1998)
34. Y Wu, C-h Chiu, Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs. IEEE Trans. Audio Speech Lang. Process. **14**(1), 266–276 (2006)
35. M Kotti, E Benetos, C Kotropoulos, Computationally efficient and robust BIC-based speaker segmentation. IEEE Trans. Audio Speech Lang. Process. **16**(5), 920–933 (2008)
36. A Misra, in *Proc. Interspeech*. Speech/Nonspeech Segmentation in Web Videos (ISCA Portland, 2012)
37. D Castán, A Ortega, A Miguel, E Lleida, Audio segmentation-by-classification approach based on factor analysis in broadcast news domain. EURASIP J. Audio Speech Music Process. **34**, 1–13 (2014)
38. A Mesa, JLN Ortega, A Teixeira, EH Pérez, PQ Morales, AR Garcia, IG Moreno, IberSPEECH 2014: VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop (2014). http://iberspeech2014.ulpgc.es/index.php/online Accessed 23 Nov 2015
39. X Anguera, J Bonastre, in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. A novel speaker binary key derived from anchor models (ISCA Makuhari, Chiba, Japan, 2010), pp. 2118–2121
40. H Delgado, C Fredouille, J Serrano, in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14–18, 2014*. Towards a complete binary key system for the speaker diarization task (ISCA Singapore, 2014), pp. 572–576
41. J Kittler, Combining classifiers: A theoretical framework. Pattern Anal. Appl. **1**(1), 18–27 (1998)

3.3. Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data

Título: "Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data"
Autores: Javier Franco-Pedroso, Daniel Ramos and Joaquin Gonzalez-Rodriguez
Revista: PLoS ONE, Volumen 11, Número 2, Febrero de 2016
Páginas: 1-25
Editor: Public Library of Science
doi: 10.1371/journal.pone.0149958
URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149958

PLOS ONE

RESEARCH ARTICLE

Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data

Javier Franco-Pedroso*, Daniel Ramos, Joaquin Gonzalez-Rodriguez

ATVS-Biometric Recognition Group, Universidad Autonoma de Madrid, Madrid, Spain

* javier.franco@uam.es

OPEN ACCESS

Citation: Franco-Pedroso J, Ramos D, Gonzalez-Rodriguez J (2016) Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. PLoS ONE 11 (2): e0149958. doi:10.1371/journal.pone.0149958

Editor: Gang Han, Taxas A&M University, UNITED STATES

Received: November 25, 2015

Accepted: January 27, 2016

Published: February 22, 2016

Data Availability Statement: Glass-fragments dataset is from the article "Aitken, C. G. G., Lucy, D. Evaluation of trace evidence in the form of multivariate data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2004;53:109–122. doi: 10.1046/j.0035-9254.2003.05271.x" and can be downloaded from: http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9876/homepage/glass-data.txt Inks and car-paints datasets are from the book "Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, Colin Aitken. Statistical 515 Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical 516 Data. Wiley; January 2014."

Abstract

In forensic science, trace evidence found at a crime scene and on suspect has to be evaluated from the measurements performed on them, usually in the form of multivariate data (for example, several chemical compound or physical characteristics). In order to assess the strength of that evidence, the likelihood ratio framework is being increasingly adopted. Several methods have been derived in order to obtain likelihood ratios directly from univariate or multivariate data by modelling both the variation appearing between observations (or features) coming from the same source (within-source variation) and that appearing between observations coming from different sources (between-source variation). In the widely used multivariate kernel likelihood-ratio, the within-source distribution is assumed to be normally distributed and constant among different sources and the between-source variation is modelled through a kernel density function (KDF). In order to better fit the observed distribution of the between-source variation, this paper presents a different approach in which a Gaussian mixture model (GMM) is used instead of a KDF. As it will be shown, this approach provides better-calibrated likelihood ratios as measured by the log-likelihood ratio cost (C_{llr}) in experiments performed on freely available forensic datasets involving different trace evidences: inks, glass fragments and car paints.

Introduction

A likelihood ratio represents a ratio of likelihoods between two competing hypothesis. In the context of forensic science, these two hypotheses are that of the prosecution, H_p (for instance, the suspect originated the crime scene mark), and that of the defence, H_d (for instance, the suspect is not the origin of the crime scene mark). If some samples of a given material coming from a known source (*control* data) and some others coming from an unknown source (*recovered* data) are given, both known as *the evidence* (E), and some other information (I) related to the crime is available, the trier of fact (judge or jury) looks for the ratio between the probabilities

and can be downloaded from: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470972106.html.

Funding: JFP recieved funding from "Ministerio de Economia y Competitividad (ES)" (http://www.mineco.gob.es/) through the project "CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz", with grant number TEC2012-37585-C02-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

of the H_p and H_d hypotheses given by

$$\frac{P(H_p|E,I)}{P(H_d|E,I)} \tag{1}$$

expressing the relative strength of one hypothesis versus the other.

However, the role of the forensic scientist must be restricted to evaluate the likelihood of the evidence assuming that any of the competing hypothesis is true, and it is not the evaluation of any other information different from that needed to evaluate the strength of the evidence. Using Bayesian theory, the above described ratio can be decomposed in the following way:

$$\frac{P(H_p|E,I)}{P(H_d|E,I)} = \frac{P(H_p|I)}{P(H_d|I)} \cdot \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(H_p|I)}{P(H_d|I)} \cdot LR \tag{2}$$

making a clear separation of the role of the forensic scientist and that of the judge or jury. Thus, the likelihood ratio (LR) strengthens ($LR > 1$) or weakens ($LR < 1$) the probabilities of the propositions, in the light of the newly observed evidence. In the process of assigning/computing the LR, additional data, usually known in forensics as *background population*, is needed to obtain the likelihood of the parameters for the model used.

A possible statement of the hypotheses at the source level [1] is:

- H_p: the samples found at the crime scene and those obtained from the suspect come from a common source.
- H_d: the samples found at the crime scene and those obtained from the suspect come from different sources.

Other forms of the hypotheses are possible [1], but the analysis is outside the scope of this paper.

Likelihood ratios can be either directly derived from the data through the application of some probabilistic models (also known as feature-based LRs) or by transforming simple raw scores from a recognition system through a calibration step [2] (also known as score-based LRs). The score-based approach has been mainly used for biometric systems [3], in which the pattern recognition process does not follow a probabilistic model but a pattern matching procedure [4], the assumed conditions does not exactly hold (e.g. observations are not i.i.d. or do not follow a normal distribution), or the number of dimensions in the feature space makes the problem intractable (e.g. image vectors [5] or GMM-means supervectors [6]). However, recent approaches in face and speaker recognition modalities have begun to apply probabilistic methods with the aid of dimensionality reduction techniques [7–9]. On the other hand, the feature-based approach is usually followed in applied statistics to forensic science [10–12], where the observations are quite stable features whose within-source variation can be modelled by a normal distribution (for instance, measurements of the concentration of some chemical compounds).

A widely used approach within forensics [12–14] is that presented in [10], where the likelihood ratio is computed from multivariate data through the application of a two-level random effect model taking into account the variation *i*) between samples coming from the same source, known as *within-source* variation, and *ii*) between samples coming from different sources, known as *between-source* variation. Within-source variation is taken to be constant and normally distributed, and expressions for both normal and non-normal distribution for the between-source variation are given. When a normal distribution can not be assumed for the between-source variation, a kernel density function (KDF) [15] is used. However, as it will

be shown, this KDF approach overestimates the between-source density function in some areas of the feature space for datasets where sources are grouped in several clusters.

In order to avoid this problem, an alternative approach is presented in this work, in which the between-source distribution is represented by means of a Gaussian mixture model (GMM) [16, 17], whose parameters are obtained through a maximum-likelihood (ML) criterion, with the aim of obtaining a better representation of how the parameter being modelled (sources mean) varies across the different sources observed in the background population. As being also a probabilistic method for clustering data, GMMs provide a better representation of such kind of datasets, which leads to obtain better calibrated likelihood ratios.

The rest of the paper is organized as follows. In Section [Likelihood ratio computation], the likelihood ratio computation method is presented and the generative model defined. Section [Models for between-source distribution] describes the expressions to be used for a normally distributed between-source variation and those to be used when it is represented by means of a Gaussian mixture; for this latter case, the KDF expression used in [10] is also shown. In Section [GMMs for non-normal between-source distributions], the GMM training process is described, and the differences between using the KDF and the GMM approaches are highlighted. Section [Experimental framework] describes the forensic databases, the experimental protocols and the evaluation metrics, while the results are presented and discussed in Section [Results and Discussion]. Finally, conclusions are drawn in Section [Conclusions].

Likelihood ratio computation

In order to compute the likelihood ratio, the probability of the evidence has to be evaluated under the two competing hypothesis, H_p and H_d, where the evidence consists in both the control ($\mathbf{y}_1$) and the recovered ($\mathbf{y}_2$) datasets (see the mathematical notation given in the [Appendix]). If H_p is assumed true, the joint probability of both datasets has to be evaluated; on the other hand, if H_d is assumed true, each dataset is generated from a different source and hence they are independent.

$$LR = \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(\mathrm{y}_1, \mathrm{y}_2|H_p)}{P(\mathrm{y}_1, \mathrm{y}_2|H_d)} = \frac{p(\mathrm{y}_1, \mathrm{y}_2)}{p(\mathrm{y}_1) \cdot p(\mathrm{y}_2)} \tag{3}$$

If a generative model with parameters Λ for the observed samples is assumed, the Bayesian solution is obtained by integrating out these parameters (if they vary from one source to another) for a given distribution which is usually obtained from a background population dataset, $p(\Lambda|\mathbf{X})$.

$$\frac{p(\mathrm{y}_1, \mathrm{y}_2)}{p(\mathrm{y}_1) \cdot p(\mathrm{y}_2)} = \frac{\int_\Lambda p(\mathrm{y}_1, \mathrm{y}_2|\Lambda)\, p(\Lambda|\mathbf{X})\, \mathrm{d}\Lambda}{\int_\Lambda p(\mathrm{y}_1|\Lambda)\, p(\mathrm{y}_2|\Lambda)\, p(\Lambda|\mathbf{X})\, \mathrm{d}\Lambda} \tag{4}$$

Final expressions for the numerator and denominator of the likelihood ratio will depend on the assumed generative model, which defines both the parameters Λ and the specific density functions. In this Section, we will describe the generative model used in [10], and the within-source distribution will be defined.

The generative model

The two-level random effect model [18] used in [10] can be seen as a generative model in which a particular observed feature vector $\mathbf{x}_{ij}$ coming from source i is generated through

$$\mathrm{x}_{ij} = \theta_i + \psi_j \tag{5}$$

where θ_i is a realization of the source random variable Θ and ψ_j is a realization of the additive random noise Ψ representing its within-source variation. This noisy term is taken to be constant among different sources and randomly distributed following

$$\Psi \sim \mathcal{N}(0, \mathbf{W}) \tag{6}$$

where $\mathbf{W}$ is the within-source covariance matrix. Thus, the conditional distribution of the random variable X_i (from which $\mathbf{x}_{ij}$ is drawn), given a particular source i, follows a normal distribution with mean θ_i and covariance matrix $\mathbf{W}$

$$\mathrm{X}_{ij}|(\Theta = \theta_i) \sim \mathcal{N}(\theta_i, \mathbf{W}) \tag{7}$$

Within-source covariance matrix can be computed from a background population dataset, comprising $N = m \cdot n$ samples coming from m different sources, through

$$\mathbf{W} = \frac{\mathbf{S}_w}{N - m} \tag{8}$$

being $\mathbf{S}_w$ the within-source scatter matrix given by

$$\mathbf{S}_w = \sum_{i=1}^{m}\sum_{j=1}^{n}(\mathrm{x}_{ij} - \bar{\mathrm{x}}_i)(\mathrm{x}_{ij} - \bar{\mathrm{x}}_i)^{\mathrm{T}} \tag{9}$$

where $\bar{\mathrm{x}}_i$ is the average of a set of n feature vectors from source i.

As the assumed generative model has only one varying parameter, θ, characterizing the particular source, and the observed samples are assumed i.i.d. conditioned on the knowledge of θ, the numerator and the denominator of the likelihood ratio given in Eq 4 can be expressed, respectively, by

$$p(\mathrm{y}_1, \mathrm{y}_2) = \int_\theta p(\mathrm{y}_1|\theta, \mathbf{W})\, p(\mathrm{y}_2|\theta, \mathbf{W})\, p(\theta|\mathbf{X})\, \mathrm{d}\theta \tag{10}$$

where the parameter θ jointly varies for both control and recovered conditional probabilities, as they are assumed to come from the same source (say $\theta_1 = \theta_2 = \theta$), and

$$p(\mathrm{y}_1) \cdot p(\mathrm{y}_2) = \int_\theta p(\mathrm{y}_1|\theta, \mathbf{W})\, p(\theta|\mathbf{X})\, \mathrm{d}\theta \times \int_\theta p(\mathrm{y}_2|\theta, \mathbf{W})\, p(\theta|\mathbf{X})\, \mathrm{d}\theta \tag{11}$$

where these conditional probabilities can be integrated out independently as they are assumed to come from different sources (say $\theta_1 \neq \theta_2$).

Similarly to the random variable X_{ij}, the conditional distribution of a random variable $\bar{\mathrm{X}}_i$ representing the average of a set of n feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n\}$ coming from a particular source i is given by

$$\bar{\mathrm{X}}_i|(\Theta = \theta_i) \sim \mathcal{N}(\theta_i, \mathbf{D}), \quad \mathbf{D} = \frac{\mathbf{W}}{n} \tag{12}$$

Thus, when evaluating the conditional probability of a set of n_1 control samples, $\mathbf{y}_1$, or a set of n_2 recovered samples, $\mathbf{y}_2$, they will be evaluated in terms of their sample mean. That is,

$$p(\mathrm{y}_l|\theta_l, \mathbf{W}) = p(\bar{\mathrm{y}}_l|\theta_l, \frac{\mathbf{W}}{n_l}) = N(\bar{\mathrm{y}}_l; \theta_l, \mathbf{D}_l), \quad l = 1, 2 \tag{13}$$

This leads to the following expressions for the previously shown integrals:

$$p(\mathrm{y}_1, \mathrm{y}_2) = \int_\theta N(\bar{\mathrm{y}}_1; \theta, \mathbf{D}_1)\, N(\bar{\mathrm{y}}_2; \theta, \mathbf{D}_2)\, p(\theta|\mathbf{X})\, \mathrm{d}\theta \tag{14}$$

and

$$p(\mathrm{y}_l) = \int_\theta N(\bar{\mathrm{y}}_l; \theta, \mathbf{D}_l)\, p(\theta|\mathbf{X})\, \mathrm{d}\theta, \quad l = 1, 2 \tag{15}$$

where only the distribution of the parameter θ remains undefined.

Models for between-source distribution

Regarding the distribution $p(\theta|\mathbf{X})$ from which the parameter characterizing the source θ is drawn, its shape depends on how the between-source variation is modelled. In this Section, two different types of distribution of such parameter, obtained from a background population, are shown. First, we will describe the expressions for a normally distributed between-source variation. While this is not the case under analysis in this work, it will serve to derive the expressions for the non-normal case, which is expressed in terms of a weighted sum of Gaussian densities.

Normal case

If sources means can be assumed normally distributed, $\boldsymbol{\Theta} \sim \mathcal{N}(\mu, \mathbf{B})$, then

$$p(\theta|\mathbf{X}) = N(\theta; \mu, \mathbf{B}) \tag{16}$$

where μ and $\mathbf{B}$ are, respectively, the mean vector and the covariance matrix of the between-source distribution. These *hyperparameters* can be obtained from a background population (with m sources, n samples per source and N total samples) through

$$\mu = \frac{1}{m}\sum_{i=1}^{m} \bar{\mathrm{x}}_i \tag{17}$$

and

$$\mathbf{B} = \frac{\mathbf{S}_b}{m-1} - \frac{\mathbf{S}_w}{n(N-m)} \tag{18}$$

where the between-source scatter matrix, $\mathbf{S}_b$, is given by

$$\mathbf{S}_b = \sum_{i=1}^{m} (\bar{\mathrm{x}}_i - \mu)(\bar{\mathrm{x}}_i - \mu)^{\mathrm{T}} \tag{19}$$

Using this distribution for the parameter θ of the generative model, the integrals involved in the likelihood ratio computation can be written

$$p(\mathrm{y}_1, \mathrm{y}_2) = \int_\theta N(\bar{\mathrm{y}}_1; \theta, \mathbf{D}_1)\, N(\bar{\mathrm{y}}_2; \theta, \mathbf{D}_2)\, N(\theta; \mu, \mathbf{B})\, \mathrm{d}\theta \tag{20}$$

and

$$p(\mathrm{y}_l) = \int_\theta N(\bar{\mathrm{y}}_l; \theta, \mathbf{D}_l)\, N(\theta; \mu, \mathbf{B})\, \mathrm{d}\theta, \quad l = 1, 2 \tag{21}$$

Using the Gaussian identities given in the Appendix, the numerator of the likelihood ratio can be shown to be equal to:

$$p(\mathrm{y}_1, \mathrm{y}_2) = N(\bar{\mathrm{y}}_1; \bar{\mathrm{y}}_2, \mathbf{D}_1 + \mathbf{D}_2) \cdot N(\bar{\mathrm{y}}^*; \mu, \mathbf{D}^* + \mathbf{B}) \tag{22}$$

where

$$\bar{\mathrm{y}}^* = (\mathbf{D}_1 + \mathbf{D}_2)^{-1}(\mathbf{D}_2\bar{\mathrm{y}}_1 + \mathbf{D}_1\bar{\mathrm{y}}_2) \tag{23}$$

and

$$\mathbf{D}^* = \mathbf{D}_1(\mathbf{D}_1 + \mathbf{D}_2)^{-1}\mathbf{D}_2 \tag{24}$$

Finally, each of the integrals in the denominator is given by

$$p(\mathrm{y}_l) = N(\bar{\mathrm{y}}_l; \mu, \mathbf{D}_l + \mathbf{B}), \quad l = 1, 2 \tag{25}$$

Non-normal case

When the normal assumption does not hold for the distribution of sources means among the background population data, the between-source distribution $p(\theta|\mathbf{X})$ can be approximated by a weighted sum of C Gaussian densities in the following form:

$$p(\theta|\mathbf{X}) = \sum_{c=1}^{C} \pi_c N(\theta; \mu_c, \Sigma_c) \tag{26}$$

where $\{\pi_k\}_{c\,=\,1,\,\ldots,\,C}$ are the weighting factors and have the following constraints

$$0 \leq \pi_c \leq 1, \quad \sum_{c=1}^{C} \pi_c = 1 \tag{27}$$

With this distribution as the prior probability for the parameter θ of the generative model, the integrals involved in the likelihood ratio computation can be written

$$\begin{aligned} p(\mathrm{y}_1, \mathrm{y}_2) &= \int_\theta \{N(\bar{\mathrm{y}}_1; \theta, \mathbf{D}_1)\, N(\bar{\mathrm{y}}_2; \theta, \mathbf{D}_2) \sum_{c=1}^{C} \pi_c N(\theta; \mu_c, \Sigma_c)\}\, \mathrm{d}\theta \\ &= \sum_{c=1}^{C} \pi_c \int_\theta \{N(\bar{\mathrm{y}}_1; \theta, \mathbf{D}_1)\, N(\bar{\mathrm{y}}_2; \theta, \mathbf{D}_2)\, N(\theta; \mu_c, \Sigma_c)\}\, \mathrm{d}\theta \end{aligned} \tag{28}$$

and

$$\begin{aligned} p(\mathrm{y}_l) &= \int_\theta \{N(\bar{\mathrm{y}}_l; \theta, \mathbf{D}_l) \sum_{c=1}^{C} \pi_c N(\theta; \mu_c, \Sigma_c)\}\, \mathrm{d}\theta \\ &= \sum_{c=1}^{C} \pi_c \int_\theta \{N(\bar{\mathrm{y}}_l; \theta, \mathbf{D}_l)\, N(\theta; \mu_c, \Sigma_c)\}\, \mathrm{d}\theta, \quad l = 1, 2 \end{aligned} \tag{29}$$

As it can be seen, the Gaussian mixture expressions become a weighted sum of the expressions given for the normal case, and so the probabilities involved in the likelihood ratio computation can be easily derived, resulting in

$$p(\mathrm{y}_1, \mathrm{y}_2) \quad = N(\bar{\mathrm{y}}_1; \bar{\mathrm{y}}_2, \mathbf{D}_1 + \mathbf{D}_2) \cdot \sum_{c=1}^{C} \pi_c N(\bar{\mathrm{y}}^*; \mu_c, \mathbf{D}^* + \Sigma_c) \tag{30}$$

and

$$p(\mathrm{y}_l) = \sum_{c=1}^{C} \pi_c N(\bar{\mathrm{y}}_l; \mu_c, \mathbf{D}_l + \Sigma_c), \quad l = 1, 2 \tag{31}$$

In [10], between-source distribution $p(\theta|\mathbf{X})$ is approximated through a KDF [15] where the kernel function $K(\cdot)$ is taken to be a multivariate normal function with smoothing parameter, or *bandwidth*, $\mathbf{H} = h^2\,\mathbf{B}$:

$$p(\theta|\mathbf{X}) = \frac{1}{m|\mathbf{H}|^{1/2}} \sum_{i=1}^{m} K\left(\frac{\theta - \bar{\mathrm{x}}_i}{\mathbf{H}^{1/2}}\right) = \frac{1}{m} \sum_{i=1}^{m} N(\theta; \bar{\mathrm{x}}_i, h^2\mathbf{B}) \tag{32}$$

where

$$h = \left(\frac{4}{2d+1}\right)^{\frac{1}{d+4}} m^{-1/(d+4)} \tag{33}$$

As it can be seen, this Gaussian KDF is in fact a Gaussian mixture whose parameters, equating terms in Eq 26, are given by

$$C = m, \;\; \pi_c = \frac{1}{m}, \;\; \mu_c = \bar{\mathrm{x}}_i, \;\; \Sigma_c = h^2\mathbf{B} \tag{34}$$

Thus, the between-source variation is approximated by an equally weighted sum of multivariate Gaussian functions placed at every source mean present in the background population, $\bar{x}_i$, being their covariance matrices given by $h^2\,\mathbf{B}$. That is, a weighted version of the between-source variation is *translated* to each source mean present in the background. As we will show later on, this will lead to overestimations of the between-source density in some areas of the feature space.

GMMs for non-normal between-source distributions

In this work, we propose to use a Gaussian Mixture Model (GMM) trained by means of a maximum-likelihood (ML) criterion in order to represent the distribution of the parameter θ characterizing the source. This model assumes that the observations are generated from a mixture of a finite number of Gaussian densities with unknown *hyperparameters*. Thus, it has been widely used to model the distribution of datasets in which the observations are grouped in several clusters, being each of them represented by a Gaussian density. In the case at hand, the observations are the means of the sources ($\bar{\mathrm{x}}_i$) present in the background population dataset ($\mathbf{X}$), from which the distribution $p(\theta|\mathbf{X})$ is going to be modelled.

GMM training

Maximum likelihood (ML) is a method of determining the parameters Φ of a model that makes the observed samples the most probable given that model. Conversely to KDF, where the parameters ($\bar{\mathrm{x}}_i$, $\mathbf{H}$) are first established and the density function $p(\theta|\mathbf{X})$ arises from them, in the GMM approach the density function is obtained by maximizing the likelihood of the observed data given the model, $p(\mathbf{X}|\Phi)$, from which the optimum parameters of the model are derived. In the case of a GMM of C components in the form of Eq 26, the ML parameters of the model, $\Phi = \{\pi_c, \mu_c, \Sigma_c\}_{c\,=\,1,\,\ldots,\,C}$, are obtained [17] by maximizing the following

log-likelihood:

$$\ln p(\mathbf{X}|\Phi) = \sum_{i=1}^{m} \ln \left\{ \sum_{c=1}^{C} \pi_c N(\bar{\mathbf{x}}_i; \mu_c, \Sigma_c) \right\} \tag{35}$$

This can be done through the well known expectation-maximization (EM) algorithm [17, 19], which is an iterative method that successively updates the parameters Φ of the model until convergence. A recipe for this iterative process can be found in [17].

For a faster convergence of the algorithm, usually some steps of the *k-means* algorithm [17, 20] are previously iterated in order to obtain a good initialization of the GMM, as this clustering method provides the mean vectors $\{\mu_c\}_{c=1,\ldots,C}$ (known as *centroids*) and the initial assignment of samples to clusters, from which $\{\pi_c\}_{c=1,\ldots,C}$ and $\{\Sigma_c\}_{c=1,\ldots,C}$ can be obtained.

The specific number of components, C, can be set by different methods. If the feature vectors are low-dimensional, the number of components can be visually estimated by inspecting a 2-D or 3-D projection of the background population data; however, depending on the structure of the data, there can be a lot of ambiguity in this process. Another option is to apply the *elbow method* [21] in the initial clustering stage, in which the cost function is plotted for different (increasing) number of clusters; for the first number of clusters there will be a great change when increasing the number of clusters, but at some point the marginal gain will drop indicating the proper number of clusters. A similar method can be applied by training GMMs for different numbers of components and evaluating the gain in terms of likelihood when increasing the number of them. Finally, similarly to the previous approach, if different GMMs for different number of components are trained, some model selection methods, like the Bayesian information criterion (BIC) [22] or the Akaike information criterion (AIC) [23], can be applied.

In this work, results are reported for several number of components in order to analyse how the evaluation metrics vary depending on this parameter, and the proper number of components related to the log-likelihood of the background data given the between-source density. For a given number of components, the *k-means* algorithm is iterated until convergence previously to the EM algorithm. In order to avoid local minima in *k-means* clustering, 100 random initializations are performed for a given number of components.

GMM versus KDF

For the purpose of illustrating the differences between KDF and GMM approaches, a synthetic 2-dimensional dataset has been generated (see Fig 1), in which 10 samples from 50 sources are drawn from normal distributions with the same covariance matrix (having then the same within-source variation). Sources means are drawn from 2 different normal distributions (25 sources each), each centred at a different separated point of the feature space, and one having a larger variance than the other in one of the dimensions. As a consequence, samples coming from different sources are grouped in two clearly separated clusters, one of them having a larger local intra-cluster between-source variation than the other. Also, the overall between-source variation is higher in one of the dimensions.

As already shown in Section [Models for between-source distribution], the density function $p(\theta|\mathbf{X})$ given by KDF approach is an equally weighted sum of Gaussian densities centred at each background source mean with covariance matrices $h^2\,\mathbf{B}$ (see Eq 32). Thus, a weighted version of the overall between-source variation is *translated* to every source mean, reproducing this variation locally at each source mean. The resulting density function $p(\theta|\mathbf{X})$ for our synthetic dataset can be seen in Fig 2, where it is shown that the local intra-cluster between-source variation in dimension 1 is highly overestimated for both clusters, and slightly overestimated in dimension 2 for one of them due to the larger variation in the other one.

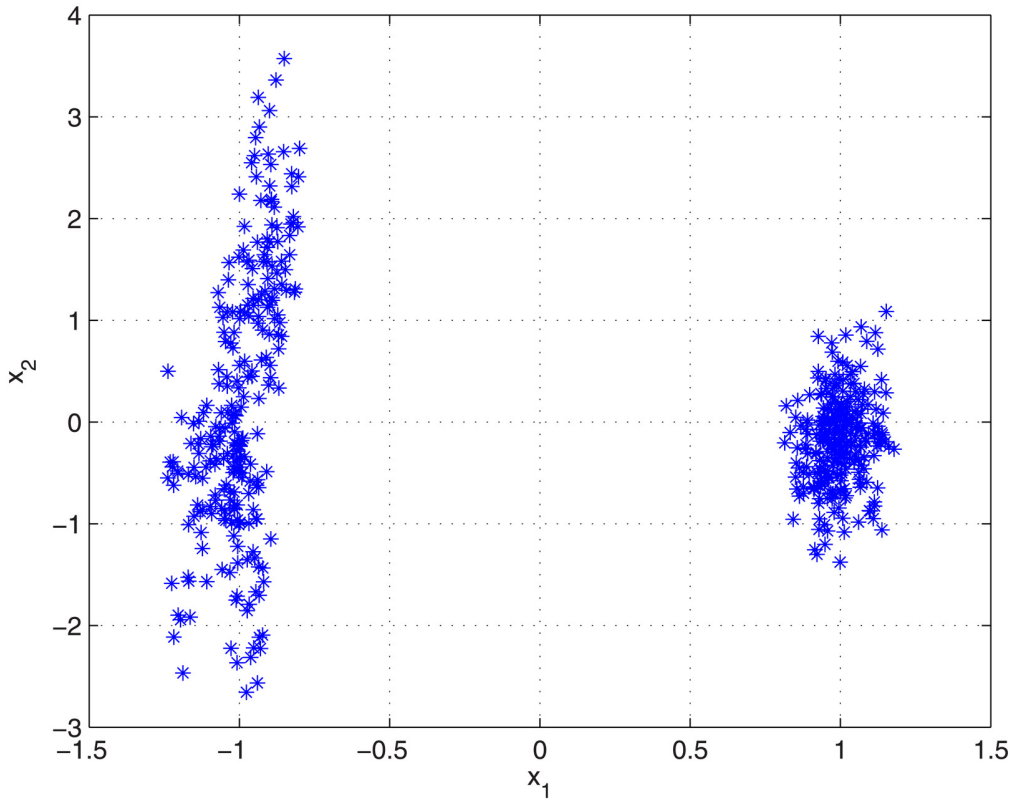


Fig 1. Synthetic dataset. Samples from a 2-dimensional synthetic dataset in which sources are grouped in two separate clusters.

doi:10.1371/journal.pone.0149958.g001

Conversely to KDF, in the GMM approach the Gaussian components are not forced to be centred at each source mean present in the background population, but a smaller number of components can be established allowing different sources means being generated from the same Gaussian component. Moreover, covariance matrices are neither fixed in advance, allowing to be locally learned for each component. As a consequence, the resulting density function can better fit the local between-source variation and the clustered nature of the dataset, as it is shown in Fig 3 for a 2-component GMM.

However, care must be taken in order to avoid *overfitting* when computing the density function through maximum likelihood. For a ML-trained GMM, the degree of fitting to the background data can be controlled through both the number of components C of the mixture and the number of EM iterations. In this work, for a given number of components, only two EM iterations are performed in order to avoid *overfitting*.

Accounting for within-source variation in the background population

When training a GMM from background sources means by maximizing the log-likelihood in Eq 35, it is assumed that there is no uncertainty in these mean values. However, the number of samples per source in the background population can be limited in forensic scenarios, and so these means cannot be reliably computed. In order to account for the uncertainty in these mean values, every observation belonging to those sources can be used to train a GMM by

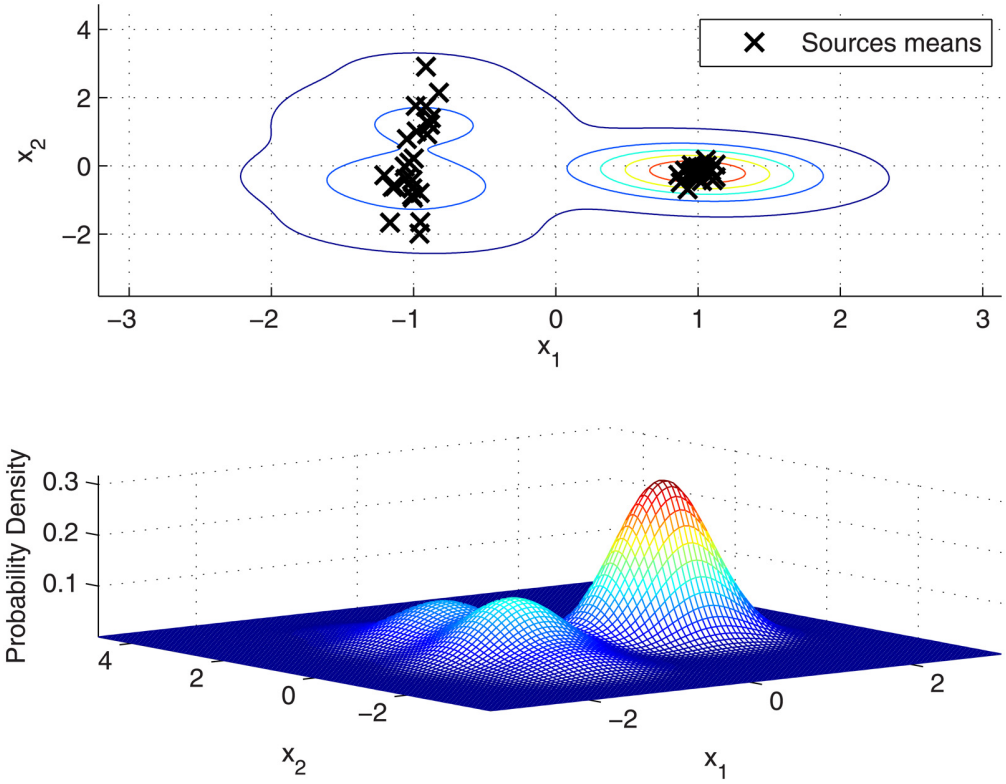


Fig 2. KDF modelling of between-source variation in the synthetic dataset. (Above) Sources means and level contours of the between-source density function. (Below) 3-dimensional representation of the between-source density function.

doi:10.1371/journal.pone.0149958.g002

maximizing the following log-likelihood:

$$\ln p(\mathbf{X}|\Phi) = \sum_{i=1}^{m}\sum_{j=1}^{n} \ln \left\{ \sum_{c=1}^{C} \pi_c N(\mathrm{x}_{ij}; \mu_c, \Sigma_c) \right\} \tag{36}$$

While there can be not much difference in the values obtained for components means μ_c in a well balanced background dataset (same number of samples per source), taking into account the variation of the samples from each source around its mean value through Eq 36 provides a more conservative background density, as every background sample is considered as a possible mean value of a source. Furthermore, this also helps to avoid Gaussian collapsing when a reduced number of sources are assigned to a particular component. The effect on our synthetic dataset is shown in Fig 4, where the Gaussian densities are placed at the same locations as in Fig 3 but larger variances and covariances are obtained, specially for the cluster with lower intra-cluster between-source variation.

Experimental framework

Forensic datasets

In order to test the approach proposed in this work, several types of forensic datasets have been used, being one of them the glass-fragments dataset also used in [10], which can be downloaded

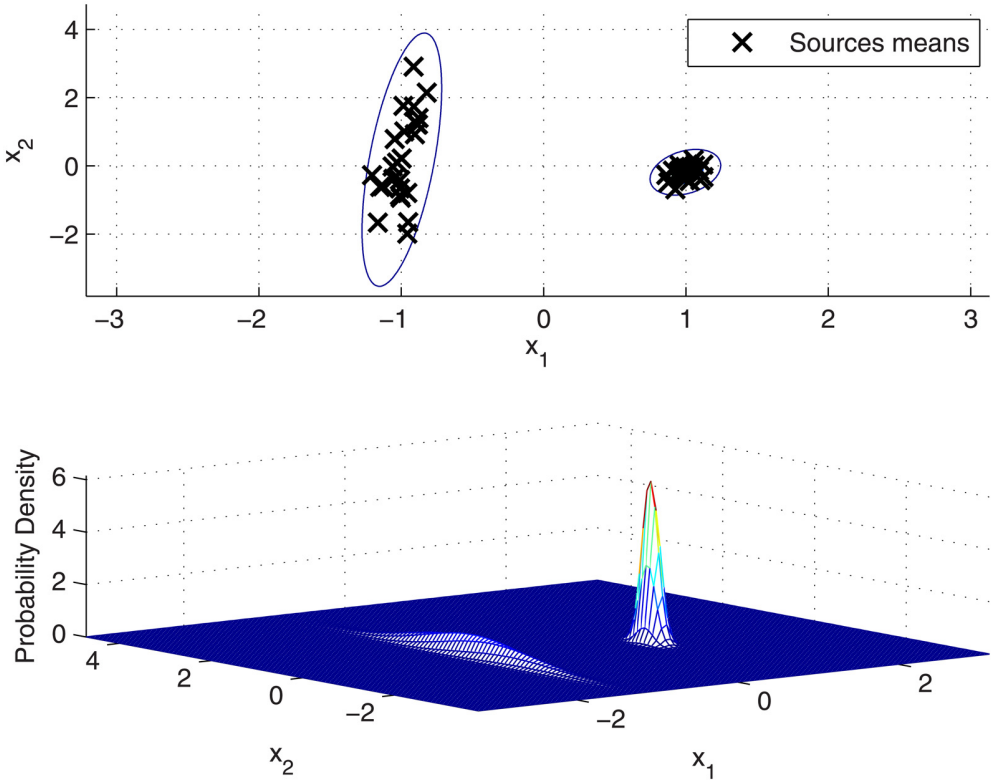


Fig 3. GMM modelling of between-source variation in the synthetic dataset. (Above) Sources means and level contours of the between-source density function. (Below) 3-dimensional representation of the between-source density function.

doi:10.1371/journal.pone.0149958.g003

from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9876/homepage/glass-data.txt. A detailed description of the other two datasets can be found in [12], and can be downloaded from http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470972106.html.

- *Inks*. For this dataset, the features are the measurements of the $d = 3$ chromaticity coordinates r, g and b (being $r + g + b = 1$) taken on samples of blue inks. The dataset comprises the measurements on $n = 10$ samples for each of the $m = 40$ different ink sources.
- *Glass fragments*. For this dataset, the features are the measurements of the concentrations in $d = 3$ elemental ratios taken on glass fragments: log(Ca/K), log(Ca/Si) and log(Ca/Fe). The dataset comprises the measurements on $n = 5$ fragments for each of the $m = 62$ different glass sources.
- *Car paints*. For this dataset, the features are the measurements of $d = 7$ organic components present in the top layer of different acrylic car paintings. The dataset comprises the measurements on $n = 3$ samples for each of the $m = 36$ different car-paint sources.

Table 1 gathers the already mentioned characteristics of these three datasets, while Figs 5, 6 and 7 show 2-dimensional projections of their sources means. As it can be seen, sources means in the last two datasets (glass fragments and car paints) present a clustered nature, while those in the first one (inks) are normally distributed [12].

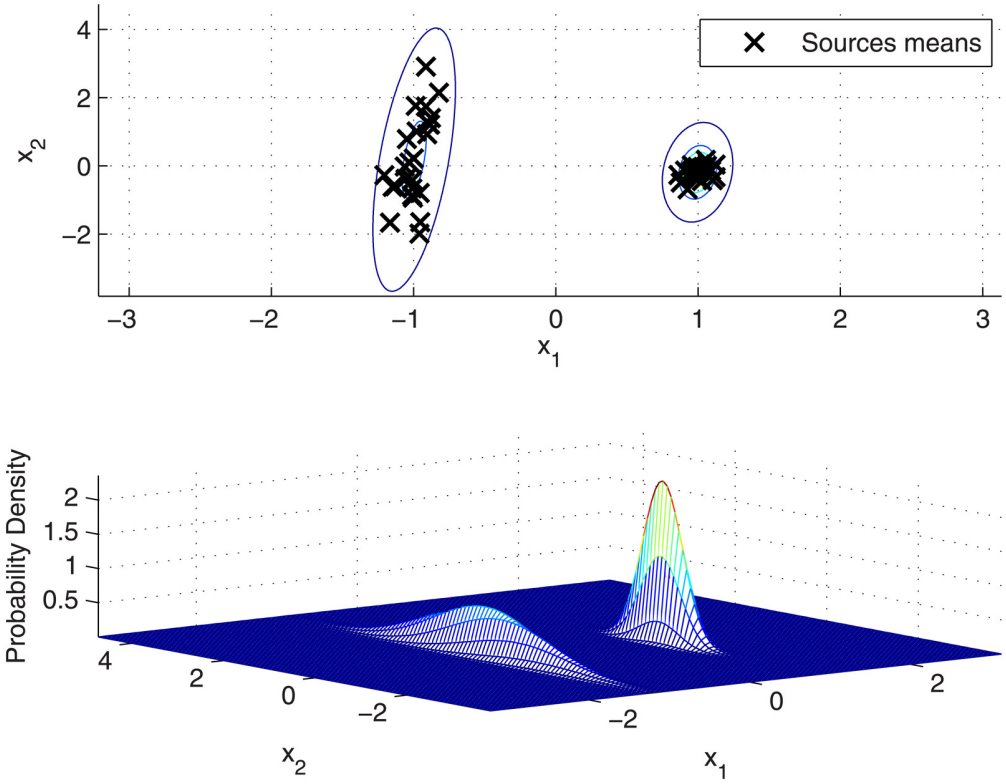


Fig 4. GMM modelling of between-source variation in the synthetic dataset when taking into account the background within-source variation. (Above) Sources means and level contours of the between-source density function. (Below) 3-dimensional representation of the between-source density function.

doi:10.1371/journal.pone.0149958.g004

Protocols

The protocol followed in [10] used the whole glass-fragment dataset in order to obtain the between-source probability density function $p(\theta|\mathbf{X})$. Then, for each source, the first 3 samples (out of 5) were used as control data and the last 3 were used as recovered data, having so both datasets one sample in common. While this *non-partitioning* protocol alleviates the lack of data due to the small size of the dataset, it may lead to overoptimistic results as the different subsets (background, control and recovered) are overlapped.

In this work, a *cross-validation* protocol is also used in order to avoid overoptimistic results, in which the dataset is divided into two non-overlapping subsets devoted to:

- obtain the between-source distribution $p(\theta|\mathbf{X})$ (*known data* or *training subset*), and

Table 1. Datasets summary.

	m	n	d
Inks	40	10	3
Glass fragments	62	5	3
Car paints	36	3	7

m, number of sources; n, number of samples per source; d, number of features.

doi:10.1371/journal.pone.0149958.t001

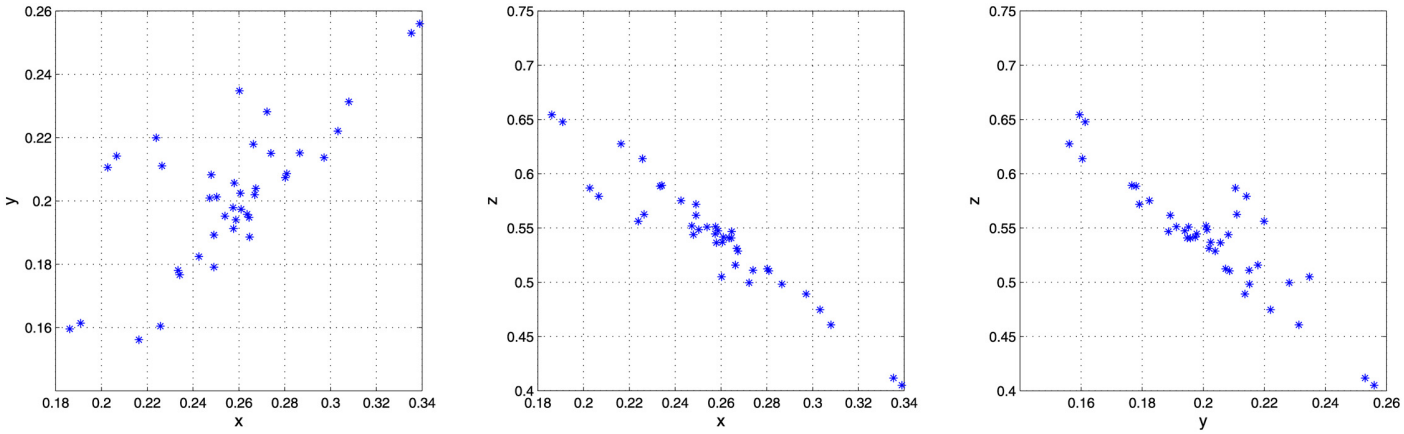


Fig 5. Sources means in the inks dataset. The three 2-dimensional projections of the sources means.

doi:10.1371/journal.pone.0149958.g005

- compute same-source and different-source likelihood ratios (*unknown data* or *testing subset*). This subset is further divided into two non-overlapping halves acting as control and recovered data.

In order to alleviate the lack of data, this procedure is carried out in the following way. For each of the $m(m-1)/2$ possible pairs of sources in the dataset, all the samples belonging to those two sources are taken apart from the dataset in order to be used as the *testing subset*, being the remaining sources used as the *training subset*. Each of the two sources in the *testing subset* is divided into two non-overlapping halves ({1a, 1b} and {2a, 2b}) that can be used either as control or recovered data to perform 2 same-source comparisons (1a-1b, 2a-2b) and 4 different-source comparisons (1a-2a, 1a-2b, 1b-2a, 1b-2b). Although the same control and recovered data from a particular source is used in all the different pairs in which it is involved, as the remaining sources change for each different pair, different between-source distributions $p(\theta|\mathbf{X})$ are involved in likelihood ratio computations. This procedure allow us to perform a total number of $m(m-1)$ same-source comparisons and $2 \times m(m-1)$ different-source comparisons for a given dataset, instead of the m same-source comparisons and $m(m-1)/2$ different-source comparisons performed in [10], while the between-source distribution $p(\theta|\mathbf{X})$ used in every

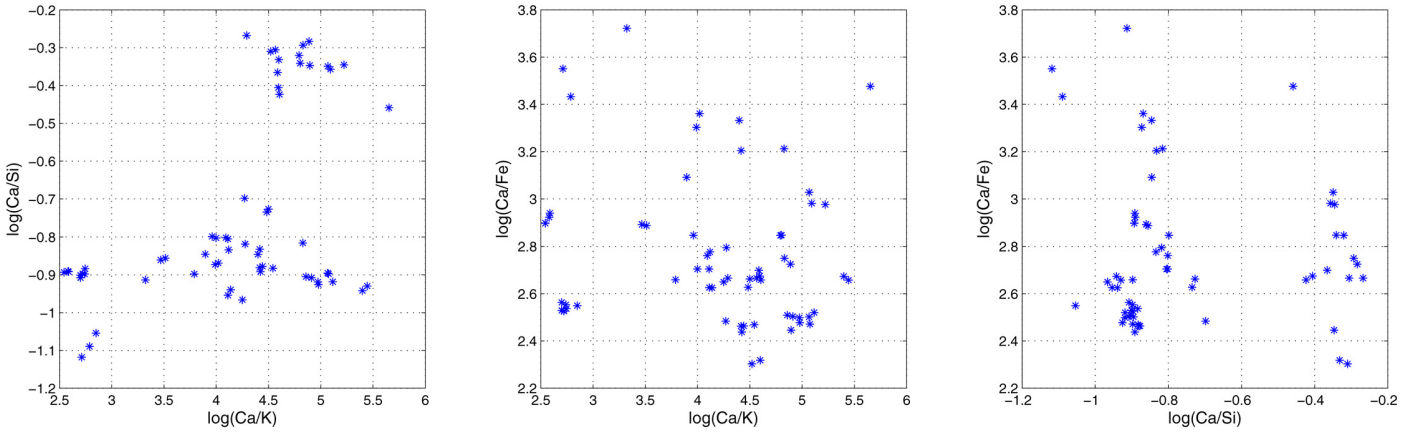


Fig 6. Sources means in the glass-fragments dataset. The three 2-dimensional projections of the sources means.

doi:10.1371/journal.pone.0149958.g006

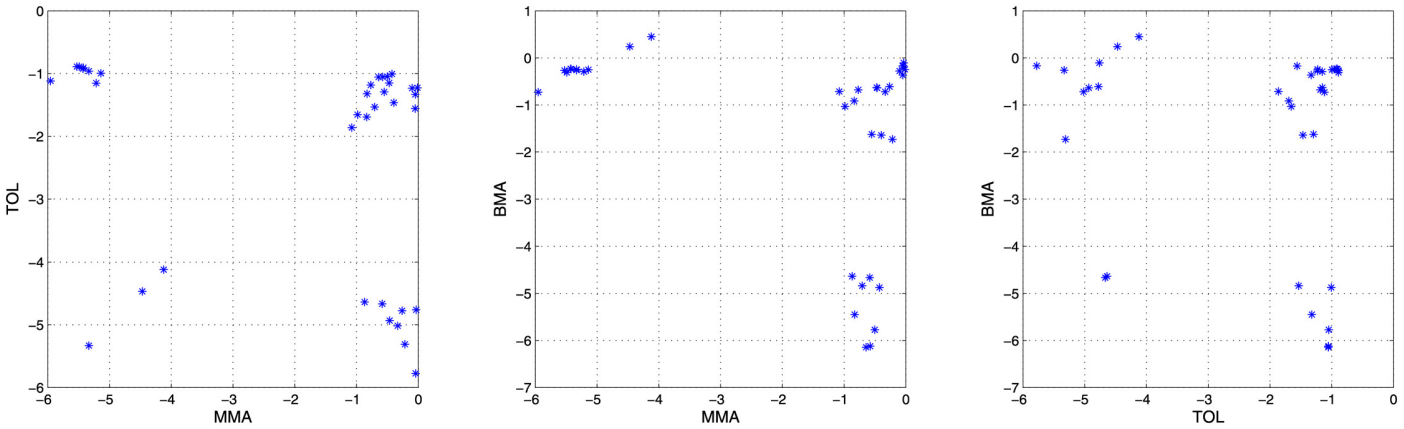


Fig 7. Sources means in the car-paints dataset. Three 2-dimensional projections of the sources means.

doi:10.1371/journal.pone.0149958.g007

comparison is obtained from $m - 2$ different sources instead of m. The specific number of comparisons for each evaluation protocol on the different datasets are given in Table 2.

Evaluation Metrics

The main evaluation metric used in order to compare the different approaches is the log-likelihood ratio cost function (C_{llr}) [2, 24], which evaluates both the *discrimination* abilities of the computed log-likelihood ratios and the goodness of their *calibration*. Given a set of log-likelihood ratios $\{\mathcal{L}\} = \{\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_C\}$ obtained from C comparisons, the C_{llr} can be computed in the following way:

$$C_{llr}(\{\mathcal{L}\}) = \frac{1}{2\log 2}\left(\frac{1}{N_{ss}}\sum_{c\in ss}\log\left(1+e^{-\mathcal{L}_c}\right)+\frac{1}{N_{ds}}\sum_{c\in ds}\log\left(1+e^{\mathcal{L}_c}\right)\right) \quad (37)$$

where '*ss*' is the set of N_{ss} same-source comparisons and '*ds*' is the set of N_{ds} different-source comparisons. As it is a cost function, the larger the C_{llr} value, the worse the verification method, being $C_{llr} = 0$ the minimum achievable cost. Note also that this metric allows to define a *neutral reference* which does not provide support for any of the two hypothesis (that is, $\mathcal{L}_c = 0$ for every comparison), providing a reference value of $C_{llr} = 1$. Thus, a verification method for which C_{llr} is larger than 1 means that it is providing misleading likelihood ratios.

An important aspect of the C_{llr} is that it can be decomposed into two additive terms, one due to the discrimination abilities (C_{llr}^{min}) and another one due to the calibration of the verification method (C_{llr}^{cal}) where

$$C_{llr}^{cal} = C_{llr} - C_{llr}^{min} \quad (38)$$

Table 2. Number of same-source and different-source comparisons in each dataset for the non-partitioning and cross-validation protocols.

	Non-partitioning		Cross-validation	
	Same-source	Different-source	Same-source	Different-source
Glass fragments	62	1891	3782	7564
Inks	40	780	1560	3120
Car paints	36	630	1260	2520

doi:10.1371/journal.pone.0149958.t002

and C_{llr}^{min} is obtained by means of the *Pool Adjacent Violators* (PAV) algorithm [25, 26] and represents the minimum achievable C_{llr} in the case of having an optimally calibrated log-likelihood ratios set $\{\mathcal{L}'\}$ (details can be found in [24]).

In order to show the performance over a wide range of prior probabilities, the Empirical Croos-Entropy (ECE) plots [27, 28] will be used. These figures (see, for example, Fig 8) graphically represent what would be the accuracy (solid curve) when using the set of logLR values $\{\mathcal{L}\}$ for each of the prior probabilities (represented as logarithmic odds) in the given range. Additionally, the discriminating power is also plotted (dashed curve) for the optimally calibrated (ideal) logLRs set $\{\mathcal{L}'\}$, along with the neutral reference (dotted curve).

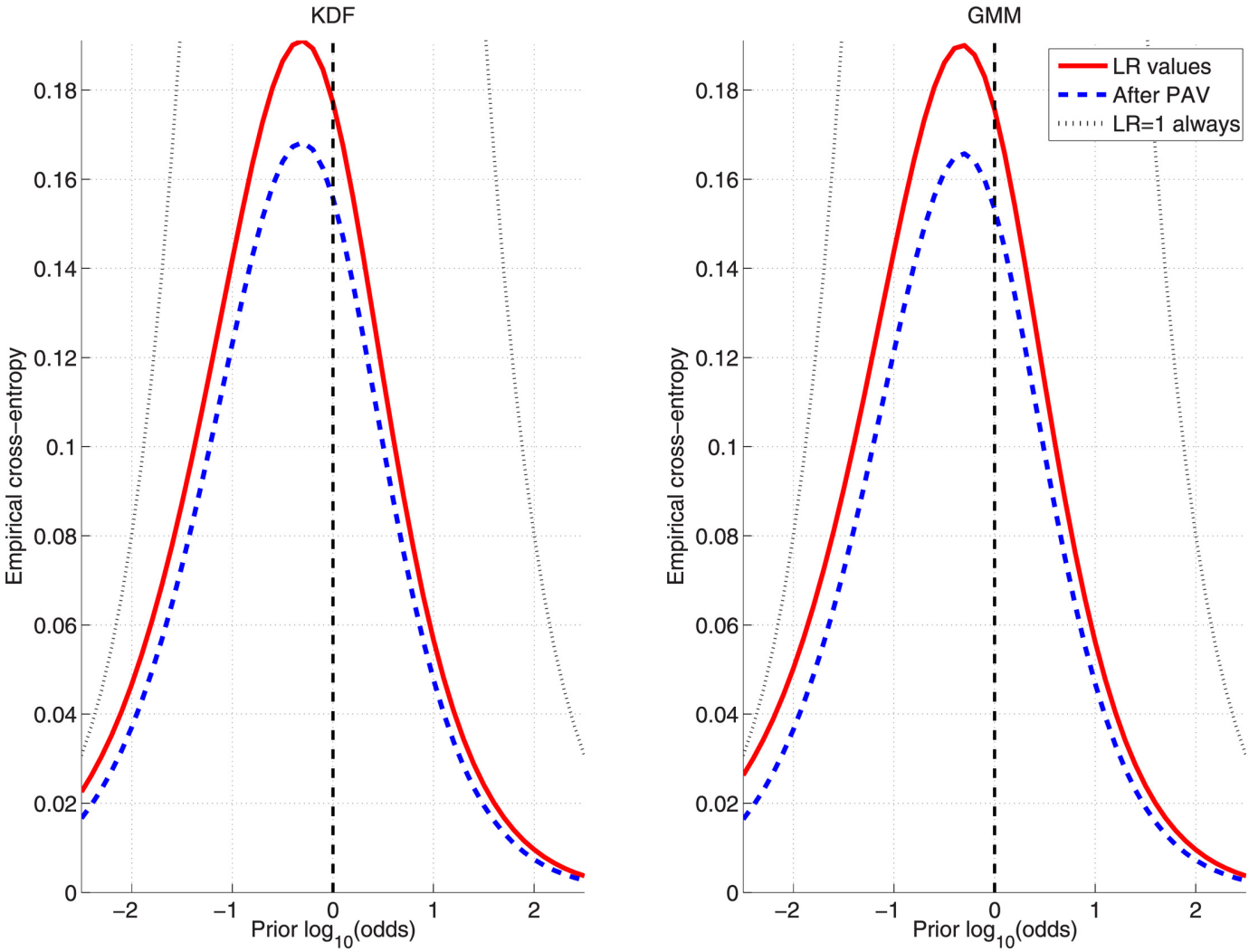


Fig 8. ECE plots for the KDF and GMM approaches on the inks dataset when applying the *cross-validation* protocol. GMM is trained by maximizing Eq 35.

doi:10.1371/journal.pone.0149958.g008

Results and Discussion

Inks dataset

For this dataset, as the background sources means are normally distributed, GMMs with a single component has been trained by maximizing either Eq 35 or Eq 36. Table 3 shows the detailed results (C_{llr}, C_{llr}^{min} and C_{llr}^{cal}) for KDF and GMM approaches (Eq 35 and Eq 36) when applying both the *non-partitioning* and the *cross-validation* protocols.

First, it should be noted that results in the *non-partitioning* protocol are slightly better for every method as it is an overoptimistic framework where data is shared between training and testing subsets. Regarding the comparison between methods, it can be seen that no significant improvement is obtained by the GMM approach as the sources means for this dataset do not present a clustered nature. Moreover, among the two GMM variants, the results obtained when maximizing Eq 35 are slightly better, presumably due to the fact that enough number of samples per source are available ($n = 10$), compared to the number of features ($d = 3$), to compute reliable sources means, and further uncertainty accounted for Eq 36 seems to be counterproductive.

Finally, Fig 8 show ECE plots for KDF and GMM (Eq 35) approaches when applying the *cross-validation* protocol, where it can be seen that both present similar performance for a wide range of prior probabilities.

Glass-fragments dataset

For this dataset, several GMMs have been trained, by maximizing Eq 35, in order to analyse how the main evaluation metric (C_{llr}) varies as a function of the number of components, C. In the experiments carried out, the maximum number of components has been limited to 6 in order to avoid Gaussian collapsing due to a reduced number of observations (sources means) per component (62 total sources in the whole dataset). Results for the *non-partitioning* protocol can be seen in Fig 9 for both KDF and GMM (Eq 35) approaches, where also the log-likelihood of the background data (sources means) given the between-source density has been plotted.

As it was expected for this *non-partitioning* protocol, C_{llr} decreases as the number of components increases, due to the shared data between training and testing subsets, which can lead to overfit the background density. However, as soon as the log-likelihood for the GMM surpass that obtained for the KDF density, better results are obtained with the GMM approach. It is also worth noting that this happens for a number of components (2–3) around that which could be expected from visual inspection of the 2-dimensional projections shown in Fig 6.

Fig 10 show the same analysis for the *cross-validation* protocol. In this case, the log-likelihood is not plotted as the GMM change for every testing sources-pair (being trained on the remaining sources). Similar conclusions than before can be drawn, but here the overfitting problem affecting the *non-partitioning* protocol is revealed, as the C_{llr} for the *cross-validation*

Table 3. Performance of KDF and GMM approaches on the inks dataset for the *non-partitioning* and *cross-validation* protocols.

	Non-partitioning			Cross-validation		
	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}
KDF	0.1459	0.0224	0.1684	0.1558	**0.0214**	0.1778
GMM (Eq 35)	**0.1430**	**0.0223**	**0.1653**	**0.1533**	0.0223	**0.1756**
GMM (Eq 36)	0.1453	0.0271	0.1724	0.1569	0.0286	0.1855

$C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$

doi:10.1371/journal.pone.0149958.t003

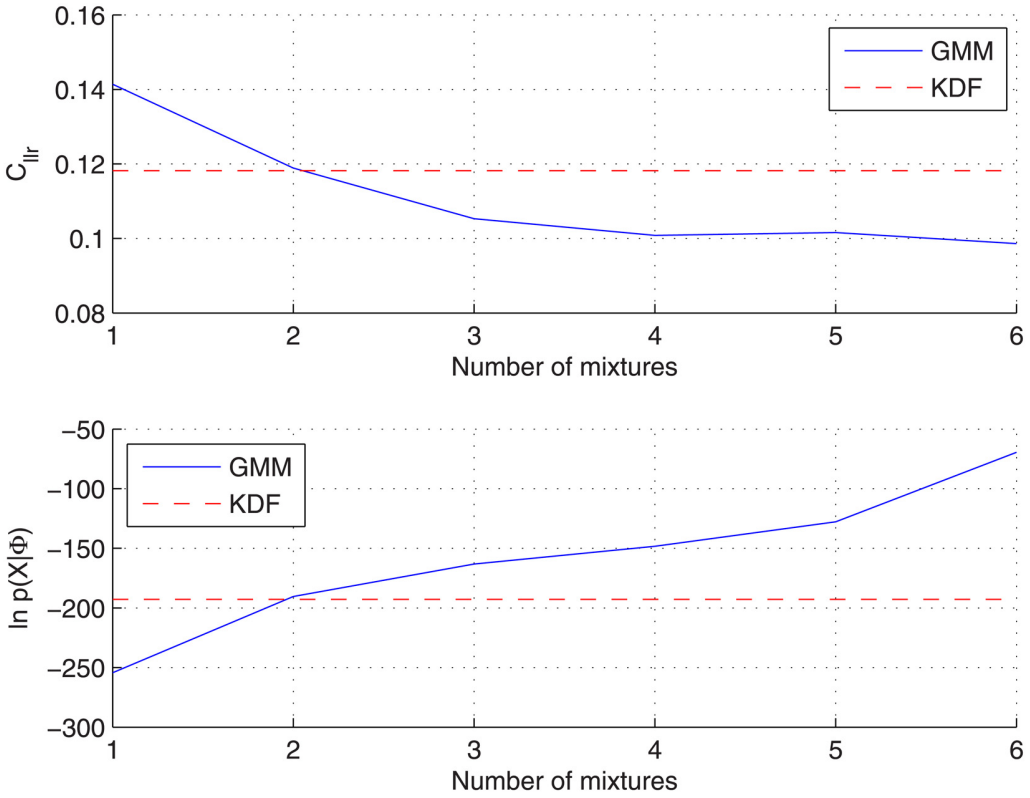


Fig 9. Analysis of the number of GMM components for the glass-fragments dataset when applying the *non-partitioning* protocol. GMM is trained by maximizing Eq 35. (Above) Log-likelihood ratio cost. (Below) Log-likelihood of the background data given the between-source density function.

doi:10.1371/journal.pone.0149958.g009

protocol reaches a minimum value for a given number of components ($C = 4$) and then increases. Results are also shown for GMMs trained by maximizing Eq 36, with similar conclusions but slightly better results, presumably due to the small number of samples per source ($n = 5$) compared to the number of features ($d = 3$).

Table 4 shows the detailed results (C_{llr}, C_{llr}^{min} and C_{llr}^{cal}) for KDF and GMM approaches (Eq 35 and Eq 36) when applying both the *non-partitioning* and the *cross-validation* protocols. For GMM approaches, results are given for the optimum number of components ($C = 4$) when the *cross-validation* protocol is applied. Again, as the *non-partitioning* protocol constitutes an over-optimistic framework, results are slightly better for every method compared to the *cross-validation* protocol. This is also the reason of obtaining better results when GMMs are trained by maximizing Eq 36, as the same sources are present in both training and testing subsets. However, when the *cross-validation* protocol is applied, there is no shared data between those subsets, and so the additional uncertainty accounted by Eq 36 provides slightly better results. In any case, both GMM approaches outperform the KDF one due to their better calibration properties for this clustered dataset.

Finally, Fig 11 shows the comparative results between KDF and GMM (Eq 36) in the form of ECE plots when the *cross-validation* protocol is applied.

Car-paints dataset

An equivalent analysis to that shown for the glass-fragments dataset has been performed for the car-paints one. Fig 12 shows both the C_{llr} and the log-likelihood of the background data

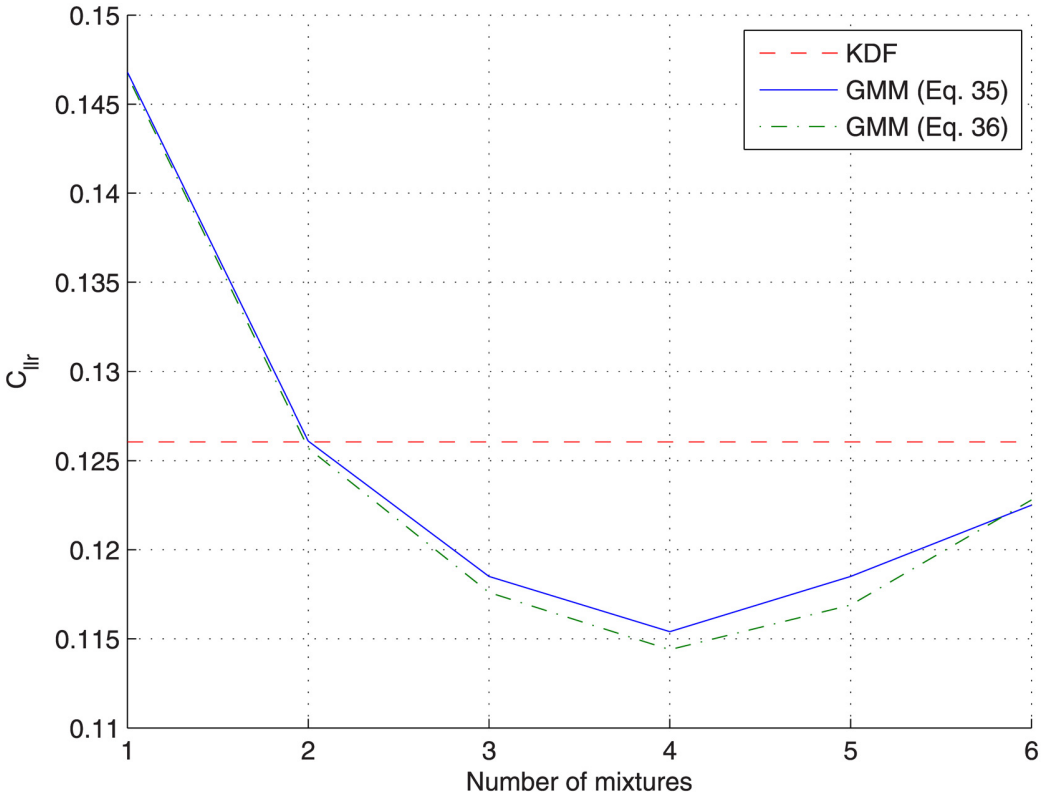


Fig 10. Analysis of the number of GMM components for the glass-fragments dataset when applying the *cross-validation* protocol.

doi:10.1371/journal.pone.0149958.g010

given the model (trained by maximizing Eq 35) as a function of the number of components for the *non-partitioning* protocol. Similarly to what happened with the previous dataset, C_{llr} decreases as the number of components increases, and as soon as the log-likelihood for the GMM surpass that obtained for the KDF density, better results are obtained with the GMM approach. Again, this happens for a number of components (3–4) around that which could be expected from visual inspection of some of the 2-dimensional projections shown in Fig 7.

Fig 13 show the same analysis for the *cross-validation* protocol (without showing the log-likelihood plot), where it can be seen (solid line) that, similarly to what happened with the glass-fragments dataset, a minimum C_{llr} value is reached for a particular number of components ($C = 3$) and then it increases. However, when plotting results for GMMs trained by maximizing Eq 36 instead (dot-dashed line), the number of components for which the minimum

Table 4. Performance of KDF and GMM approaches on the glass-fragments dataset for the *non-partitioning* and *cross-validation* protocols.

	Non-partitioning			Cross-validation		
	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}
KDF	0.0787	0.0394	0.1182	0.0850	0.0410	0.1260
GMM (Eq 35)	**0.0785**	**0.0223**	**0.1008**	0.0863	0.0291	0.1154
GMM (Eq 36)	**0.0785**	0.0229	0.1013	**0.0862**	**0.0282**	**0.1144**

$C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$

doi:10.1371/journal.pone.0149958.t004

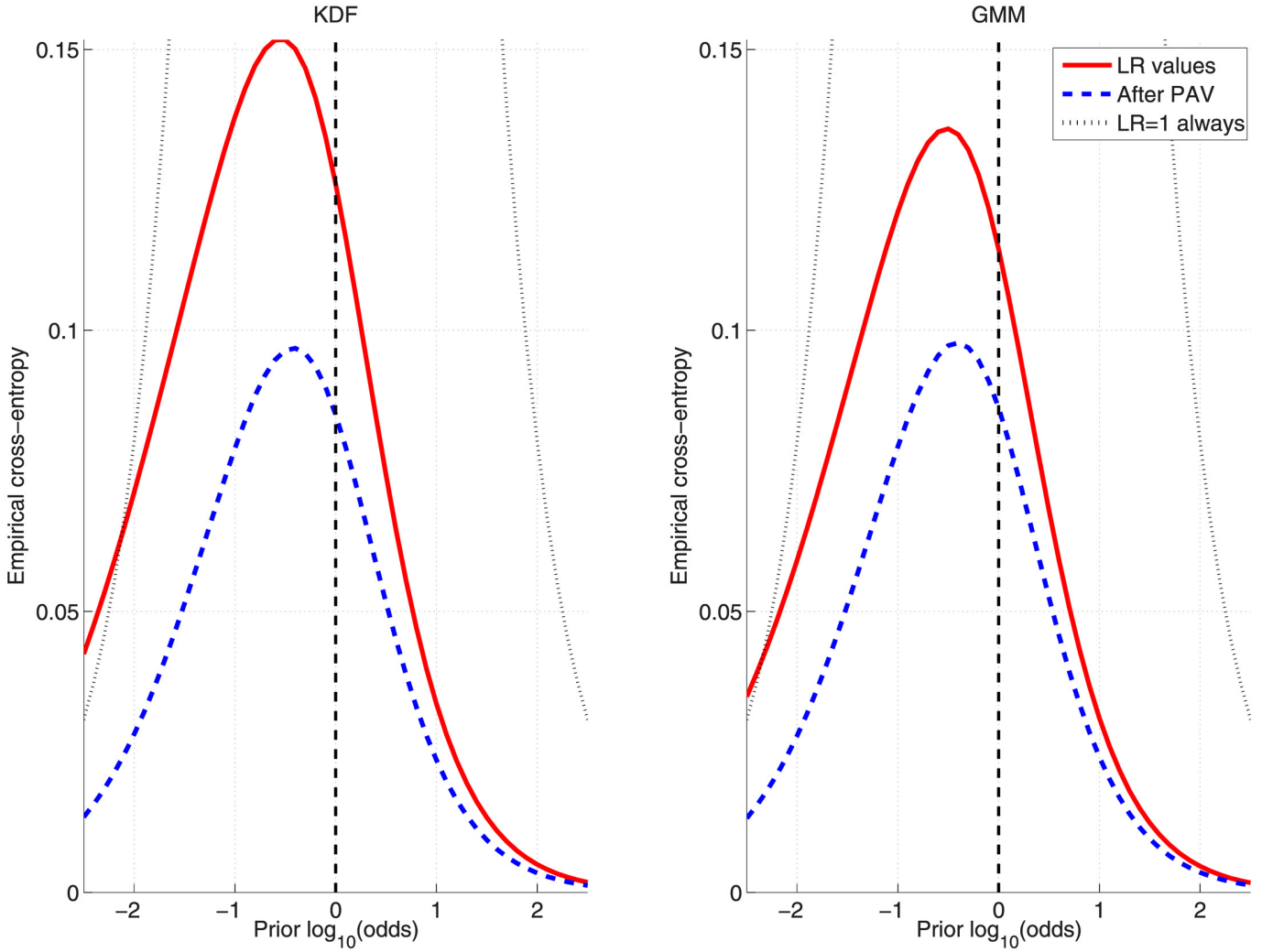


Fig 11. ECE plots for the KDF and GMM approaches on the glass-fragments dataset when applying the *cross-validation* protocol. GMM is trained by maximizing Eq 36.

doi:10.1371/journal.pone.0149958.g011

C_{llr} value is reached is slightly larger ($C = 5$); this also happens for the *non-partition* protocol, as the log-likelihood of the training data (observations) given the model for the GMM do not surpass that of the KDF until a larger number of components ($C = 4$) is reached.

Table 5 shows the detailed results (C_{llr}, C_{llr}^{min} and C_{llr}^{cal}) for KDF and GMM approaches (Eq 35 and Eq 36) when applying both the *non-partitioning* and the *cross-validation* protocols. For GMM approaches, results are given for the optimum number of components ($C = 4$ for Eq 35, $C = 5$ for Eq 36) when the *cross-validation* protocol is applied. Similar conclusions to those obtained for the glass-fragments dataset can be drawn, but much better results are obtained by GMMs approaches presumably due to the distance among clusters, which lead to KDF densities which overestimate the between-source distribution in some areas of the feature space (as shown in Fig 2 for the synthetic dataset). Among GMM approaches, the maximization of Eq 36 leads to much better results for the *cross-validation* protocol due to the small number of samples per source ($n = 3$) compared to the number of features ($d = 7$), which lead to unreliably computed sources means when training GMMs by maximizing Eq 35.

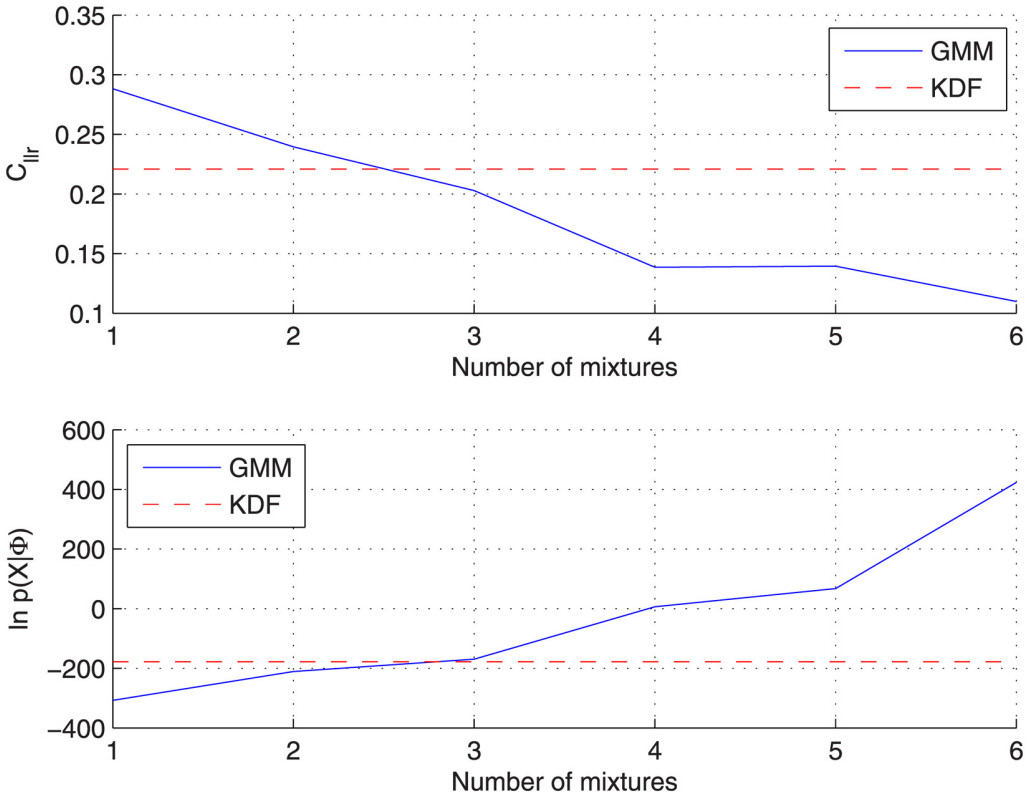


Fig 12. Analysis of the number of GMM components for the car-paints dataset when applying the *non-partitioning* protocol. GMM is trained by maximizing Eq 35. (Above) Log-likelihood ratio cost. (Below) Log-likelihood of the background data given the between-source density function.

doi:10.1371/journal.pone.0149958.g012

Finally, Fig 14 shows the comparative results between KDF and GMM (Eq 36) in the form of ECE plots when the *cross-validation* protocol is applied.

Conclusions

In this work, we present a new approach for computing likelihood ratios from multivariate data in which the between-source distribution is obtained through ML training of the parameters of a GMM. Using the same generative model as in [10], a common derivation of the LR expressions is presented for both Gaussian KDF and GMM, in which the between-source distribution is represented in terms of a weighted sum of Gaussian densities. Then, differences between KDF and GMM approaches are highlighted, and the effects on the obtained probability density are shown for a synthetic dataset. Furthermore, a variant in GMM training has been tested in order to account for the uncertainty in sources means when few samples per source are available in the background data.

The proposed approach has been tested on three different forensic datasets and compared with the KDF approach. Additionally to the *non-partitioning* protocol applied in [10], a more realistic *cross-validation* protocol is applied in order to avoid overoptimistic results, as ML-trained GMMs can overfit the background population density. Performance is evaluated in terms of the log-likelihood ratio cost function (C_{llr}), which allows to decompose the performance in a term due to the discrimination abilities and another one due to the calibration

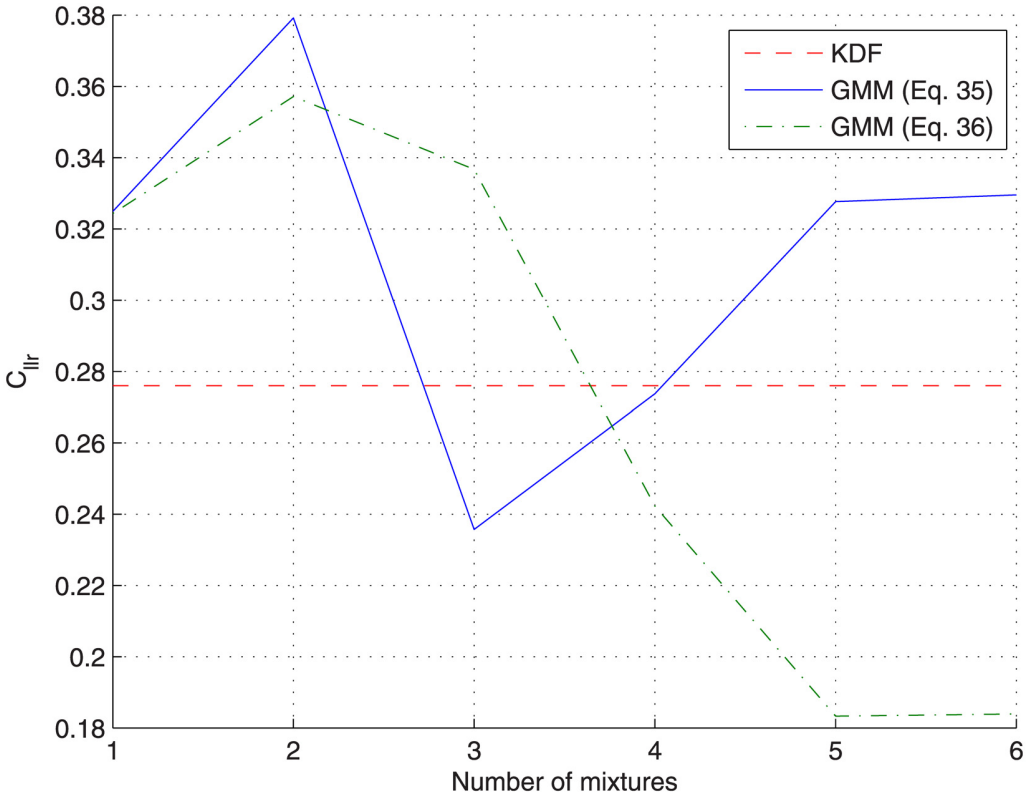


Fig 13. Analysis of the number of GMM components for the car-paints dataset when applying the *cross-validation* protocol.

doi:10.1371/journal.pone.0149958.g013

properties. ECE plots have been used to show the behaviour in a wide range of prior probabilities, which is needed in forensic science.

Results show that, although KDF and GMM approaches present similar discrimination abilities, when the datasets have a *clustered* nature, the between-source distribution is better described by a GMM, leading to better calibrated likelihood ratios. If clusters are not easily distinguishable, the between-source distribution still can be modelled by one single component, obtaining similar results to the KDF approach. Specially remarkable are the results obtained for the car-paints dataset, where $\sim$50% improvement in terms of calibration performance is obtained.

Table 5. Performance of KDF and GMM approaches on the car-paints dataset for the *non-partitioning* and *cross-validation* protocols.

	Non-partitioning			Cross-validation		
	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}
KDF	0.0819	0.1388	0.2208	0.0972	0.1786	0.2759
GMM (Eq 35)	**0.0715**	**0.0671**	**0.1386**	0.0968	0.1769	0.2737
GMM (Eq 36)	**0.0715**	0.0729	0.1443	**0.0899**	**0.0934**	**0.1833**

$C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$

doi:10.1371/journal.pone.0149958.t005

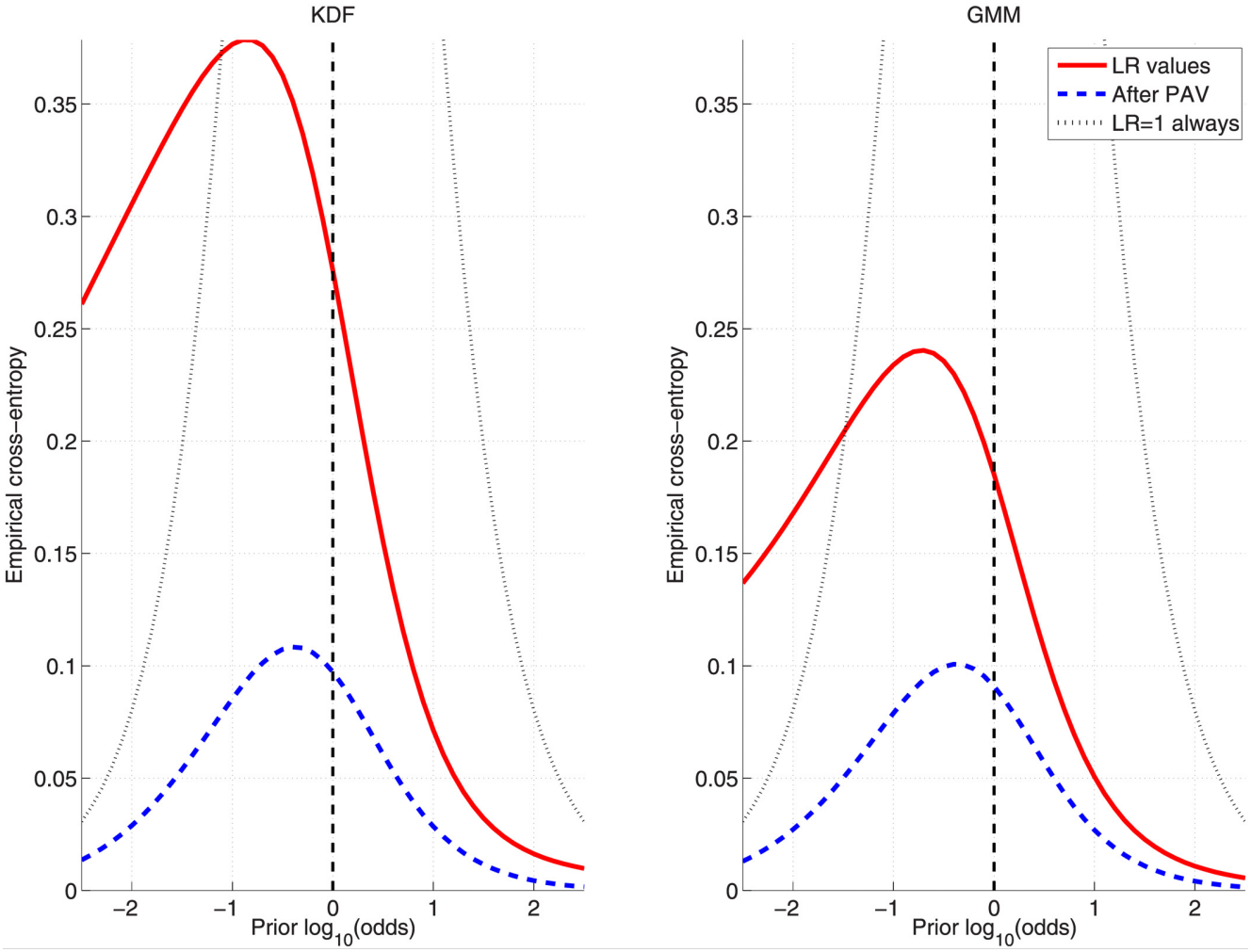


Fig 14. ECE plots for the KDF and GMM approaches on the car-paints dataset when applying the *cross-validation* protocol. GMM is trained by maximizing Eq 36.

doi:10.1371/journal.pone.0149958.g014

Appendix

Mathematical notation

Throughout this work we consider multivariate data in the form of d-dimensional column vectors $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. Following the same notation as in [10], a set of n elements of such data belonging to the same particular source i are denoted by $\mathbf{x}_i = \{\mathbf{x}_{ij}\}_{j=1,..,n} = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in}\}$, while their sample mean is denoted by $\bar{\mathbf{x}}_i$. Similarly, $\mathbf{x}_i$ is used to denote background data while $\mathbf{y}_l$ is used to denote either control ($\mathbf{y}_1$) or recovered data ($\mathbf{y}_2$). The set of feature vectors coming from different sources present in the background data is denoted by $\mathbf{X}$.

In general, column vectors are denoted by bold lower-case letters and matrices by bold upper-case letters, while scalar quantities are denoted by lower-case italic letters. Random variables are denoted by upper-case non-italic letters. $P(\cdot)$ is used to indicate the probability of a certain event, while $p(\cdot)$ denotes a probability density function. We denote a d-dimensional Gaussian distribution with mean μ and covariance matrix Σ by $\mathcal{N}(\mu, \Sigma)$ and the corresponding probability density function by $N(\mathbf{x};\mu, \Sigma)$ ($\mathrm{x} \in \mathbb{R}^d$).

Multivariate Gaussian function

$$N(\mathrm{x};\mu,\Sigma)=(2\pi)^{-d/2}|\Sigma|-1/2\ \exp\left\{-\frac{1}{2}(\mathrm{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathrm{x}-\mu)\right\}=N(\mu;\mathrm{x},\Sigma) \tag{39}$$

Gaussian identities

Product of two multivariate Gaussian functions.

$$N(\mathbf{x};\mathbf{a},\mathbf{A})\cdot N(\mathbf{x};\mathbf{b},\mathbf{B})=N(\mathbf{a};\mathbf{b},\mathbf{a}+\mathbf{B})\cdot N(\mathbf{x};\mathbf{c},\mathbf{C}) \tag{40}$$

$$\mathbf{c}=\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a}+\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} \tag{41}$$

$$\mathbf{C}=\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} \tag{42}$$

Convolution of two multivariate Gaussian functions.

$$\int_{\mathrm{x}}N(\mathbf{x};\mathbf{a},\mathbf{A})\,N(\mathbf{y}-\mathbf{x};\mathbf{b},\mathbf{B})\,\mathrm{d}\mathbf{x}=N(\mathbf{y};\mathbf{a}+\mathbf{b},\mathbf{A}+\mathbf{B}) \tag{43}$$

Expressions for a normal between-source distribution

Derivation of the numerator. First, we solve the product of the two Gaussian functions depending on either the control or the recovered data means, obtaining the following expression

$$\begin{aligned}p(\mathrm{y}_1,\mathrm{y}_2) &= \int_\theta\{N(\bar{\mathrm{y}}_1;\theta,\mathbf{D}_1)\,N(\bar{\mathrm{y}}_2;\theta,\mathbf{D}_2)\,N(\theta;\mu,\mathbf{B})\}\,\mathrm{d}\theta\\ &= \int_\theta\{N(\theta;\bar{\mathrm{y}}_1,\mathbf{D}_1)\,N(\theta;\bar{\mathrm{y}}_2,\mathbf{D}_2)\,N(\theta;\mu,\mathbf{B})\}\,\mathrm{d}\theta\\ &= \int_\theta\{N(\theta;\mathbf{z},\mathbf{Z})\,N(\bar{\mathrm{y}}_1;\bar{\mathrm{y}}_2,\mathbf{D}_1+\mathbf{D}_2)\,N(\theta;\mu,\mathbf{B})\}\,\mathrm{d}\theta\end{aligned} \tag{44}$$

where

$$\mathbf{z}=(\mathbf{D}_1+\mathbf{D}_2)^{-1}(\mathbf{D}_2\bar{\mathrm{y}}_1+\mathbf{D}_1\bar{\mathrm{y}}_2) \tag{45}$$

and

$$\mathbf{Z}=\mathbf{D}_1(\mathbf{D}_1+\mathbf{D}_2)^{-1}\mathbf{D}_2 \tag{46}$$

Being $N(\bar{\mathrm{y}}_1;\bar{\mathrm{y}}_2,\mathbf{D}_1+\mathbf{D}_2)$ independent of θ, we can solve the remaining integral as a convolution of two Gaussian functions:

$$\begin{aligned}p(\mathrm{y}_1,\mathrm{y}_2) &= N(\bar{\mathrm{y}}_1;\bar{\mathrm{y}}_2,\mathbf{D}_1+\mathbf{D}_2)\int_\theta\{N(\mathbf{z};\theta,\mathbf{Z})\,N(\theta;\mu,\mathbf{B})\}\,\mathrm{d}\theta\\ &= N(\bar{\mathrm{y}}_1;\bar{\mathrm{y}}_2,\mathbf{D}_1+\mathbf{D}_2)\int_\theta\{N(\mathbf{z}-\theta;0,\mathbf{Z})\,N(\theta;\mu,\mathbf{B})\}\,\mathrm{d}\theta\\ &= N(\bar{\mathrm{y}}_1;\bar{\mathrm{y}}_2,\mathbf{D}_1+\mathbf{D}_2)\cdot N(\mathbf{z};\mu,\mathbf{Z}+\mathbf{B})\end{aligned} \tag{47}$$

Finally, replacing $\mathbf{D}_l = \mathbf{W}/n_l$, $l = 1, 2$, in $\mathbf{z}$ and $\mathbf{Z}$

$$\mathbf{z} = \left(\frac{\mathbf{W}}{n_1} + \frac{\mathbf{W}}{n_2}\right)^{-1}\left(\frac{\mathbf{W}}{n_2}\bar{\mathrm{y}}_1 + \frac{\mathbf{W}}{n_1}\bar{\mathrm{y}}_2\right) = \frac{\frac{1}{n_2}\bar{\mathrm{y}}_1 + \frac{1}{n_1}\bar{\mathrm{y}}_2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{n_1\bar{\mathrm{y}}_1 + n_2\bar{\mathrm{y}}_2}{n_2 + n_1} = \bar{\mathrm{y}}^* \tag{48}$$

$$\mathbf{Z} = \frac{\mathbf{W}}{n_1}\left(\frac{\mathbf{W}}{n_1} + \frac{\mathbf{W}}{n_2}\right)^{-1}\frac{\mathbf{W}}{n_2} = \frac{\frac{\mathbf{W}}{n_1 n_2}}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{\mathbf{W}}{n_2 + n_1} \tag{49}$$

we obtain

$$p(\mathrm{y}_1, \mathrm{y}_2) = N(\bar{\mathrm{y}}_1; \bar{\mathrm{y}}_2, \frac{\mathbf{W}}{n_1} + \frac{\mathbf{W}}{n_2}) \cdot N(\bar{\mathrm{y}}^*; \mu, \frac{\mathbf{W}}{n_1 + n_2} + \mathbf{B}) \tag{50}$$

Derivation of the denominator. Each of the integrals in the denominator of the LR can be solved by the convolution of two Gaussian functions

$$\begin{aligned} p(\mathrm{y}_l) &= \int_\theta \{N(\bar{\mathrm{y}}_l; \theta, \mathbf{D}_l)\, N(\theta; \mu, \mathbf{B})\}\, \mathrm{d}\theta \\ &= \int_\theta \{N(\bar{\mathrm{y}}_l - \theta; 0, \mathbf{D}_l)\, N(\theta; \mu, \mathbf{B})\}\, \mathrm{d}\theta \\ &= N(\bar{\mathrm{y}}_l; \mu, \mathbf{D}_l + \mathbf{B}) = N(\bar{\mathrm{y}}_l; \mu, \frac{\mathbf{W}}{n_l} + \mathbf{B}) \end{aligned} \tag{51}$$

giving the following final expression for the denominator of the LR under the between-source normal assumption:

$$p(\mathrm{y}_1) \cdot P(\mathrm{y}_2) = N(\bar{\mathrm{y}}_1; \mu, \frac{\mathbf{W}}{n_1} + \mathbf{B}) \cdot N(\bar{\mathrm{y}}_2; \mu, \frac{\mathbf{W}}{n_2} + \mathbf{B}) \tag{52}$$

Author Contributions

Conceived and designed the experiments: JFP DR JGR. Performed the experiments: JFP. Analyzed the data: JFP DR. Contributed reagents/materials/analysis tools: JFP DR. Wrote the paper: JFP DR JGR.

References

1. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: deciding which level to address in casework. Science and Justice. 1998; 38(4):231–239. doi: 10.1016/S1355-0306(98)72117-3
2. van Leeuwen DA, Brümmer N. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. Speaker Classification I: Lecture Notes in Computer Science.2007; 4343:330–353. doi: 10.1007/978-3-540-74200-5_19
3. Gonzalez-Rodriguez J, Rose P, Ramos D, Toledano DT, Ortega-Garcia J. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. IEEE Transactions on Audio, Speech, and Language Processing. 2007; 15(7):2104–2115. doi: 10.1109/TASL.2007.902747
4. Jain AK, Ross A, Pankanti S. Biometrics: a tool for information security. IEEE Transactions on Information Forensics and Security. 2006; 1(2):125–143. doi: 10.1109/TIFS.2006.873653
5. Turk MA, Pentland AP. Face recognition using eigenfaces. Proceedings of the 1991 CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1991;586–591.

6. Campbell WM, Sturim DE, Reynolds DA. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters. 2006; 13(5):308–311. doi: 10.1109/LSP.2006.870086
7. Li P, Fu Y, Mohammed U, Elder JH, Prince SJD. Probabilistic Models for Inference about Identity. IEEE Transactions on Pattern Analysis and Machine Intelligence. January 2012; 34(1):144–157.
8. Sizov A, Lee KA, Kinnunen T. Unifying Probabilistic Linear Discriminant Analysis Variants in Biometric Authentication. Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science. 2014; 8621:464–475.
9. Borgstrom BJ, McCree A. Supervector Bayesian speaker comparison. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013;7693–7697.
10. Aitken CGG, Lucy D. Evaluation of trace evidence in the form of multivariate data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2004; 53:109–122. doi: 10.1046/j.0035-9254.2003.05271.x
11. Aitken CGG, Taroni F. Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd Edition. Wiley; July 2004.
12. Zadora G, Martyna A, Ramos D, Aitken C. Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data. Wiley; January 2014.
13. Rose P. Forensic Voice Comparison with Monophthongal Formant Trajectories—a likelihood ratio-based discrimination of "Schwa" vowel acoustics in a close social group of young Australian females. Proceedings of the 40th ICASSP International Conference on Acoustics, Speech and Signal Processing. 2015;4819–4823.
14. Bolck A, Weyermann C, Dujourdy L, Esseiva P, van den Berg J. Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. Forensic Science International. 2009; 191(1–3):42–51. doi: 10.1016/j.forsciint.2009.06.006 PMID: 19608360
15. Epanechnikov VA. Non-Parametric Estimation of a Multivariate Probability Density. Theory Probab. Appl., 14(1):153–158. doi: 10.1137/1114019
16. McLachlan GJ, Basford KE. Mixture models: Inference and applications to clustering. Applied Statistics. 1988.
17. Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.
18. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. Biometrics. 1982; 38(4):963–974. doi: 10.2307/2529876 PMID: 7168798
19. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39(1):1–38.
20. MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. 1967;281–297.
21. Ketchen DJ, Shook CL. The application of cluster analysis in Strategic Management Research: An analysis and critique. Strategic Management Journal. 1996; 17(6):441–458. doi: 10.1002/(SICI)1097-0266(199606)17:6%3C441::AID-SMJ819%3E3.0.CO;2-G
22. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978; 6(2):461–464. doi: 10.1214/aos/1176344136
23. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19(6):716–723. doi: 10.1109/TAC.1974.1100705
24. Brümmer N, du Preez J. Application-independent evaluation of speaker detection. Computer Speech and Language. 2006; 20:230–275. doi: 10.1016/j.csl.2005.08.001
25. Ahuja RK, Orlin JB. A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. Operations Research. 2001; 49:784–789. doi: 10.1287/opre.49.5.784.10601
26. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002;694–699.
27. Ramos D, Gonzalez-Rodriguez J, Zadora G, Aitken C. Information-Theoretical Assessment of the Performance of Likelihood Ratio Models. Journal of Forensic Sciences. November 2013; 58(6):1503–1518.
28. Ramos D, Gonzalez-Rodriguez J. Reliable support: measuring calibration of likelihood ratios. Forensic Science International. 2013; 230:156–169. doi: 10.1016/j.forsciint.2013.04.014 PMID: 23664798

3.4. Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009

Título: "Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009"
Autores: Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo Torre Toledano and Joaquin Gonzalez-Rodriguez
Revista: IEEE Journal of Selected Topics in Signal Processing
Volumen 4, Issue 6, Septiembre de 2010
Páginas: 1084-1093
Editor: IEEE
doi: 10.1109/JSTSP.2010.2076071

Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, *Member, IEEE*, Doroteo T. Toledano, *Member, IEEE* Joaquin Gonzalez-Rodriguez, *Member, IEEE*

***Abstract*—This work presents the systems submitted by the ATVS Biometric Recognition Group to the 2009 Language Recognition Evaluation (LRE'09), organized by NIST. New challenges included in this LRE edition can be summarized by three main differences with respect to past evaluations. Firstly, the number of languages to be recognized expanded to 23 languages from 14 in 2007, and 7 in 2005. Secondly, the data variability has been increased by including telephone speech excerpts extracted from Voice of America (VOA) radio broadcasts through Internet in addition to Conversational Telephone Speech (CTS). The third difference was the volume of data, involving in this evaluation up to 2 terabytes of speech data for development, which is an order of magnitude greater than past evaluations. LRE'09 thus required participants to develop robust systems able not only to successfully face the session variability problem but also to do it with reasonable computational resources. ATVS participation consisted of state-of-the-art acoustic and high-level systems focussing on these issues. Furthermore, the problem of finding a proper combination and calibration of the information obtained at different levels of the speech signal was widely explored in this submission. In this work, two original contributions were developed. The first contribution was applying a session variability compensation scheme based on Factor Analysis (FA) within the statistics domain into a SVM-supervector (SVM-SV) approach. The second contribution was the employment of a novel backend based on anchor models in order to fuse individual systems prior to one-vs-all calibration via logistic regression. Results both in development and evaluation corpora show the robustness and excellent performance of the submitted systems, exemplified by our system ranked 2^{nd} in the 30 second open-set condition, with remarkably scarce computational resources.**

***Index Terms*—Language Recognition, Factor Analysis, Sufficient Statistics, Linear Scoring, Anchor Models, Calibration.**

I. Introduction

RECENTLY, Spoken Language Recognition (SLR) has experienced an increase in interest mainly due to its use in a wide range of applications such as audio indexing, information retrieval or call center monitoring. While interest in the field has been latent for nearly 40 years [1], it has not been up to the last decade when systems have experienced a major research development. Among the driving factors of this rapid development and performance improvement of state-of-the-art technologies, the efforts of the US National Institute of Standards and Technologies (NIST) deserve special mention [2]. The Language Recognition Evaluations (LRE), organized by NIST since 1996, with editions in years 1996, 2003, 2005, 2007 and 2009 have established a common framework for the development and assessment of language recognition technology, succesfully focusing the efforts of the scientific community in the field. This framework includes common protocols and databases for experimental evaluation as well as well-defined evaluation methodologies [2]. Currently, the LRE evaluation has become the major and reference forum for scientific researchers and technology developers in the area who aim at adapting their systems to real-world challenges. Following such objectives, the ATVS Biometric Recognition Group of the Universidad Autonoma de Madrid (hereafter, ATVS) has been participating in LRE's since 2005, submitting systems at both lower (spectral) and higher levels (phonotactic, prosodic) for blind and public competition. From the perspective of the scientific community, the problem of automatic SLR represents a very attractive task for several reasons. On the one hand, in order to yield good performance, different levels of information across the speech signal have to be exploited. This fact implies the use of efficient methods to combine complementary information extracted from the speech signal. This is one of the major challenges in the field and it is an underlying theme in this paper. Moreover, SLR systems share most of the problems with other related research areas such as speech and speaker recognition and therefore similar solutions can be ported across to each of these fields. A good example is the inter-session variability problem, understood as the set of acoustic differences between utterances, which are not related respectively to the speaker or language to recognize. In fact, this problem, caused by several variability sources (such as channel conditions or environmental noise), is still a major source of system performance degradations in all recognition disciplines involving speech signals [3]. Because of its configuration, the LRE'09 edition clearly focused on these challenges. Session variability is present in the task by including telephone speech from Voice Of America (VOA)[1], a vast multilanguage data source new to those evaluations in addition to well-known Conversational Telephone Speech (CTS). In addition to this, a larger number of languages (23) were included, involving more language pairs difficult to distinguish (e.g Dari-Farsi, Hindi-Urdu, Bosnian-Croatian). Moreover, a huge amount of data was available to develop the systems, which required to proccess a much larger quantity of trials with respect to other evaluations. This fact highlighted the importance of systems with an acceptable balance between recognition performance and computational resources. The aim of

[1] http://www.voanews.com/english/index.cfm

this article is to describe the systems submitted by ATVS to LRE'09, which were focused on these new challenges as well as to explain some original contributions which were incorporated. The ATVS submission consisted of four different combinations of acoustic and phonotactic subsystems. The two ATVS spectral (also known as acoustic) subsystems were based in session variability compensated first-order sufficient statistics via Factor Analysis (FA) [4][5][6][7]. These statistics were calculated in our primary acoustic system which is based on the FA-GMM linear scoring framework [4], also outlined in this work as being a critical part of our acoustic systems. A novel approach, using a SVM supervector [8] acoustic system feeded from session variability compensated first-order statistics is included. The phonotactic components were based on PhoneSVM [9] composed of seven ATVS tokenizers and three tokenizers made available by Brno University of Technology (BUT). System combination is performed in a front-end-back-end configurarion. The front-end consists of recognizers trained for different languages for each of the systems used in the submission. In particular, 22 recognizers trained with VOA speech and 14 CTS recognizers trained with CTS speech were used for each system. Each recognizer for each system yielded a score, and all scores together formed a vector. After that, a back-end stage was used for classifying the resulting vector for each target language. A contribution of our submission was the use of a novel Anchor-Model approach for back-end fusion, where score vectors were classified using an SVM. Front-end scores were channel-dependent (22 VOA/14 CTS) t-normalized [10] while back-end scores are channel-independent (23 VOA+CTS) t-normalized. Calibration was achieved by the use of linear, two-class logistic regression [11], where scores were transformed into two-class, one-vs.-all log-likelihood-ratios (log-LR). In this way, a score can be interpreted as a degree of support towards any of the relevant hypotheses in the recognition process, namely $_0$ (the language in the utterance is the target language) and $_1$ (the language in the utterance is not the target language) [12]. This also allows to use Bayes thresholds for decision making, which are independent of the distribution of the output scores. The same logLR sets were submitted to the closed- and open-set conditions of the evaluation.

This paper is organized as follows. First, the ATVS individual spectral and high-level systems are described in Sections II and III respectively. Section IV presents the fusion scheme and calibration carried out in order to obtain final submitted scores, while Section V details the experimental framework for both, development and evaluation assessment. Section VI presents the ATVS submitted systems and notes on implementation details. Achieved results are presented in Section VII. Finally, future work and conclusion are outlined in Section VIII.

II. ATVS Spectral Systems

A. *FA-GMM Linear Scoring System*

The ATVS Factor Analysis Linear Scoring GMM system (hereafter, FA-GMM-LS) is based on the work developed by Niko Brummer in [4]. This system establishes a robust and efficient generative GMM framework where data sufficient statistics, relative to an Universal Background Model (UBM), play a central role. Indeed, once these are computed, both features and UBM can been discarded for next steps, with the corresponding computational savings. The *linear* term refers to novel scoring approach based on a linear approximation to log-likelihood ratios via first-order Taylor series [13]. Thus, scoring procedure simplifies to a single vector dot product. Further, session variability compensation via Factor Analysis (FA) [14] [7] is applied directly at the statistics level in both train and test stages. This subsection gives an overview of this system in four steps, where foundations for the original contributions presented in II-B are established.

*1) **Sufficient statistics**:* Given a utterance, with set of features $O = \{o_1, o_2, ..., o_n\}$ in ^{D}, and a reference model $_{UBM} = \{w_k, \mu_k, \ _k\}$, $k = 1, ...C$, zero and first-order Baum-Welch statistics, for gaussian k of $_{UBM}$, are defined as follows:

$$zero{-}order\ statistic - \quad n_k = \sum_t P_{kt} \tag{1}$$

$$first{-}order\ statistic - \quad x_k = \sum_t P_{kt} o_t \tag{2}$$

where *Gaussian Occupation Probability* P_{kt} is given by:

$$P_{kt} = P(k|o_t, \ _{UBM}) = \frac{w_k p_k(o_t)}{\sum_{j=1}^{C} w_j p_j(o_t)} \tag{3}$$

being:

$$p_k(x) = \frac{1}{(2\ )^{\frac{D}{2}} \ ^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k) \ \ ^{-1}_k (x-\mu_k)) \tag{4}$$

For convenience first-order statistics x_k use to be measured relative to the means of the model:

$$x_{norm,k} = \sum_t P_{kt}(o_t - \mu_k) \tag{5}$$

Hereafter, we refer as $\overline{x}$ to first-order statistics supervector built as the concatenation of all $x_{norm,k}$ and N as the $CDxCD$ diagonal matrix built as C blocks defined as $N_k = n_k I$, being I the DxD identity matrix.

*2) **Classical MAP**:* As in classical GMM-UBM framework [15], a GMM for each language is derived via Maximum a Posteriori Estimation (MAP) [16] from the UBM and available training data. However, here, only means are adapted and this is performed via a single MAP iteration. This shortcut besides the linear scoring approach allow to calculate only once sufficient statistics from the data and make independent the rest of the system with respect to the UBM.
In terms of sufficient statistics, MAP process to obtain a new means of a language model $_L$ can be resumed as the following equation in matrix form:

$$\overline{\mu}_L = \overline{\mu}_{UBM} + (\ I + N)^{-1} \overline{x} \tag{6}$$

where is the relevance factor and N, $\overline{x}$ resumes available training data for language L. Note that second order statistics are not neccessary because variances are not adapted.

*3) **Session variability via Factor Analysis at statistic level**:* Session variability subspace adaptation in model domain can be also seen as a mean adaptation restricted to a subspace [5][7] in the form:

$$\overline{\mu}_L = \overline{\mu}_{UBM} + Uz \tag{7}$$

where U is a low rank matrix whose columns define the session variability subspace, and z are the *channel factors*. Given U and assuming that z is normal distributed $N(O, I)$, it can be shown that finding a point estimate of z which maximizes (7) can be done by solving:

$$z = A^{-1}b = A^{-1}U \quad {}^{-1}\overline{x} \tag{8}$$

where:

$$A = I + U \quad {}^{-1}NU \tag{9}$$

$$b = U \quad {}^{-1}\overline{x} \tag{10}$$

(N.B. adapting only means, $\quad = \quad_{UBM}$)

However, it is desirable to apply the compensation in a stage before rather than in model domain as this would allow applying the compensation to test data without the need to create a model. In order to apply channel compensation directly in the statistics domain, the work in [6] where channel compensation is applied in the feature domain will serve as inspiration.
In [6], *channel* compensation is applied in every feature of an utterance i as follows:

$$\hat{o_t^{(i)}} = o_t^{(i)} - \sum_k P_{kt}U_k z \tag{11}$$

This idea can be reused in statistics domain in order to get a channel-compensated first-order statistic $\overline{y}$, in the following way:

$$\overline{y} = \overline{x} - Uz = \overline{x} - NUA^{-1}U \quad {}^{-1}\overline{x} \tag{12}$$

This approach has the desirable property of avoiding the need of a computational expensive *frame by frame* compensation.

*4) **Classical scoring vs linear scoring**:* Classical GMM-UBM scoring of a dataset X and a target model $\quad_L$ is presented as a likelihood ratio as:

$$score_{X,\ L} = \frac{P(X|\ _L)}{P(X|\ _{UBM})} \tag{13}$$

taking logarithms for practical issues this simplifies to:

$$score_{X,\ L} = log(P(X|\ _L)) - log(P(X|\ _{UBM})) \tag{14}$$

Linear scoring proposes a linear approximation of $log(P(X|\ _L))$ based on its first-order Taylor's series expansion evaluated in $\overline{\mu}_{UBM}$:

$$\begin{aligned} \log P(X|\ _L) \qquad & \log P(X|\ _{UBM}) + \\ & + \ _\mu \log P(X|\ _{UBM})\ (\overline{\mu} - \overline{\mu}_{UBM}) \end{aligned} \tag{15}$$

Several advantages of this approach with respect to classical scoring, arise by carefully analizing the above Equation (15). First of all, the need to compute term $logP(X|\ _{UBM})$ is removed, being cancelled as easily shown substituting Equation (15) into Equation (14) as follows:

$$\begin{aligned} score_{X,\ L} &= \log P(X|\ _{UBM}) + \ _\mu \log p(X|\ _{UBM}) \\ & \quad (\overline{\mu} - \overline{\mu}_{UBM}) - log(P(X|\ _{UBM})) \\ &= \ _\mu \log p(X|\ _{UBM})\ (\overline{\mu} - \overline{\mu}_{UBM}) \end{aligned} \tag{16}$$

Further, term $(\overline{\mu} - \overline{\mu}_{UBM})$ is just the offset in a classical MAP adaptation in which only a EM iteration is done. Taking advantage of this fact, target models can be expressed in FA-GMM-LS as the *offsets* in MAP adaptation, $m = (\ I+N)^{-1}\overline{x}$ (see Equation (6)), since the need of using a UBM is removed from this step on.

Moreover, it can be shown that term $\ _\mu \log p(X|\ _{UBM})$ is the first-order statistics $\overline{x}$ but normalized by the diagonal covariance matrix [17]. Thus, the scoring function is reduced to a dot product between the MAP offset model m and the first-order statistics calculated from X with respect to the UBM and normalized by the diagonal covariances matrix.

Summarizing the previous analysis, the score between a model $\quad_L$ generated from sufficient statistics N_{train} and $\overline{x}_{train}$ and a test dataset X represented by its first-order statistic $\overline{x}_{test}$ is defined by:

$$\begin{aligned} score_{X,\ L} &= (\overline{\mu} - \overline{\mu}_{UBM}) \cdot (\ {}^{-1}\overline{x}_{test}) \\ &= (\ I + N_{train})^{-1}\overline{x}_{train} \cdot (\ {}^{-1}\overline{x}_{test}) \end{aligned} \tag{17}$$

Note that in order to apply session variability compensation in both train and test phases, first order statistics $\overline{x}_{train}$ and $\overline{x}_{test}$ must be replaced by compensated stats $\overline{y}_{train}$ and $\overline{y}_{test}$ following Equation 12.

B. SVM Working on Session Variability Compensated Supervectors

The ATVS SVM supervector (SVM-SV) system is based on the work proposed in [18] where a GMM mean supervector is considered a point in the high-dimensional transformed space where the SVM works. Each GMM mean supervector represents a mapping between an utterance and a high-dimensional vector and thus, the need for explicitly performing a mapping from a lower dimensional space as in GLDS approach [8] is avoided. Then, an hyperplane is estimated in this SVM subspace to discriminatively separate a target class from non-target classes.

A modification to the work in [18] was introduced into our system by employing a session variability compensation scheme within the statistics domain, by using the channel compensated first-order statistics from the FA-GMM-LS system. Then, a single MAP adaptation was applied in order to obtain compensated GMM supervectors.

Even though others channel compensated techniques applied to SVM have been proposed in the literature [19][6][20], as far as author's knowledge, none of them have been designed to work at this level, where its application implies some advantages. On one hand, although session variability compensation techniques applied to the feature domain such as feature Nuissance Attribute Projection (fNAP) [21] or

feature Latent Factor Analysis (fLFA) [6][21] have the prime advantage of allowing any type of posterior modeling, its application implies a frame-by-frame compensation over the set of features rather than a single compensation in model or statistics domain. This becomes a major drawback when large amounts of data must be processed, as in language recognition. On the other hand, once first-order statistics are channel compensated, no other FA techniques applied at model domain such as [20] or NAP [19] were necessary. This turned out in a major saving of computational time in our acoustic systems as well as a significant benefits in terms of recognition performance.

III. High Level Systems

Even though the ATVS submissions to recent LRE's have also included a prosodic system, in LRE'09 all our high-level systems were based on phonotactic systems. Among high-level systems, phonotactic systems are one of the most successful and classic approaches in the field of language recognition [22]. Phonotactic systems try to model the sequences of phonemes that are characteristic of a particular language by processing speech with a Phonetic Recognizer (PR) that transforms speech into a sequence of phonetic tokens. Systems can use a single PR or many different PRs in different languages (Parallel PR, or PPR) for better performance. The set of languages of the PRs does not need to meet with those to be recognized, which is highly desirable because otherwise it would be necessary to train a new PR for each new language to recognize.

The sequence of recognized phonetic tokens can be used in different ways for language recognition. The most classical approach is to use statistical Language Modelling (LM) techniques to model the frequencies of phones and phone sequences (n-grams) for each particular language. The combination of a single PR and LM gives the Phone Recognition Language Modelling (PRLM) approach [22]. The language model (LM_i) is previously trained on the phonetic sequences obtained by the PR from utterances known to be of language i. It is common to use also a Universal Background Model with a structure similar to the language models but trained on phonetic sequences obtained from many languages to represent the generality of all languages through a PR. Once these two models are available, the first step to verify the language of the utterance is to process it with the PR to produce the phonetic sequence, X. Then, the phonetic decoding of the test utterance, X, and the statistical models (LM_i, UBM) are used to compute the likelihoods of the phonetic decoding, X, given the language model LM_i and the background model UBM. The recognition score is the log of the ratio of both likelihoods, normalized by the number of phonemes in the phonetic sequence. Global scheme of this process is shown in Figure 1. As different PRs can be used for the same task, it is common to use a combination of several PRs and LM in an approach known as Parallel-PRLM (PPRLM) [22]. This approach dominated the field of language recognition for years and is still, with some evolutions and improvements, one key subsystem of state-of-the-art language recognition systems.

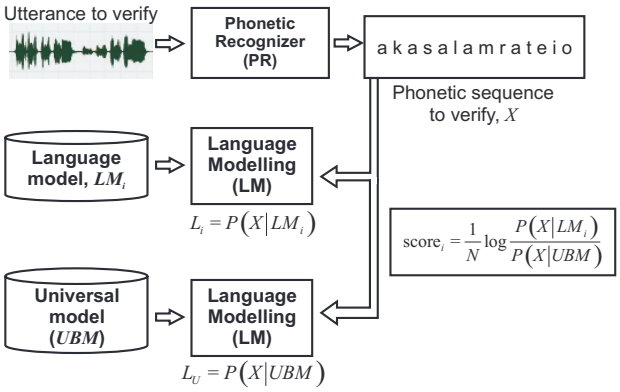


Fig. 1: *Verification process in PRLM scheme.*

One of the most important recent improvements in terms of performance is the use of SVMs for classifying the whole n-gram probability matrices [9], instead of using them in a likelihood ratio framework. This last type of system is usually referred to as Phone-SVM and is the type of system used in ATVS submission to LRE'09.

IV. Fusion and Calibration

As previously stated, a complete language recognition system is usually a combination of many individual subsystems. Combining this information by efficiently using the complementary information of every subsystem involved is known as fusion. The back-end/fusion strategy presented in this work and used in the LRE'09 evaluation is based on the use of an anchor models scheme [23].

Recently, the anchor models approach has been successfully used for both speaker verification and language identification [24][25] but not with the goal of fusion. The idea behind this approach is not only modelling the distribution of the scores for a target language with the scores for every utterance belonging to this language but to take advantage of the distribution of these scores against non-target models as well. By using anchor models, each utterance is mapped into a model space, called anchor model space, where the relative behaviour of the speech utterance with respect to other models can be learned. A point in this space is built by simply stacking scores obtained for testing an utterance over the cohort of pre-trained model as shown in Figure 2 (a). Once the set of stacked scores vectors are obtained for each language, these are used as inputs of a SVM system for discriminative purposes. Incorporating new subsystems to this fusion scheme is trivial as can be shown in Figure 2 (b).

In order to take the actual detection decision we have followed a per-language detection approach to calibrate the output log-likelihood-ratios (log-LR). Each score for each of the 23 target languages in the evaluation has been mapped to a logLR assuming a target-language-vs-rest configuration (one-vs-all). Therefore, each score can be interpreted as follows:

$$s_{cal} = \log(LR) = \log \frac{p(s|\ 0)}{p(s|\ 1)} \tag{18}$$

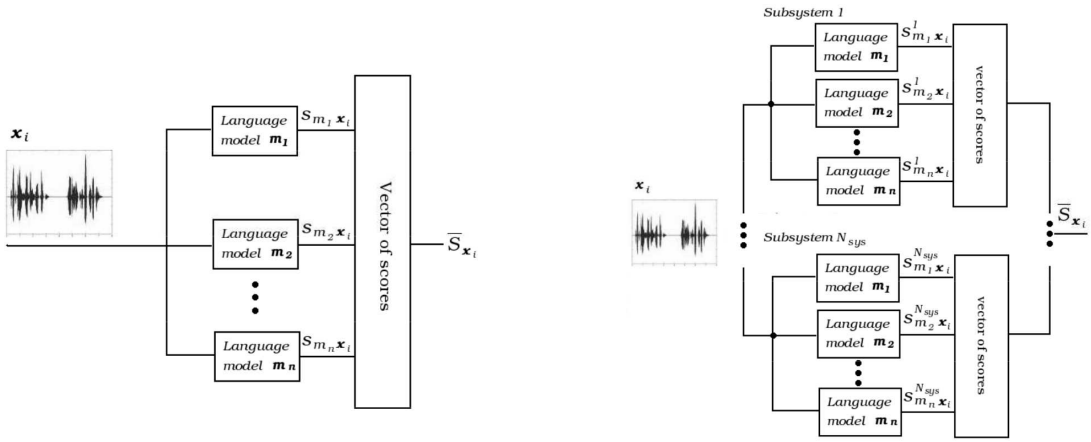


Fig. 2: *a) n-class parallel language detection problem where $\overline{S}_{x_i}$ stacks the similarities of x_i (input signal) over the set of models m_j, b) Generation of features (scores) in the anchor model space.*

where s_{cal} is the calibrated score, s is the score to be calibrated, and the hypotheses are defined as follows:

- $_0$: the language in the test utterance is the target language.
- $_1$: the language in the test utterance is not the target language.

Thus, a different score-to-log-LR mapping is performed per target language, and therefore the calibration strategy has been conducted independently for each target language. Linear logistic regression [11] has been trained, using the FoCal toolkit [2], on the complete development set of scores for each language.

After calibrating log-LR values, the logarithm of the Bayes threshold has been used in order to take decisions, defined as:

$$\log(\ _B) = \log \frac{P_{nt} C_{fa}}{P_t C_{fr}} \tag{19}$$

where $P_{nt} = P_t = 0.5$ and $C_{fa} = C_{fr} = 1$ as defined by NIST, and therefore $\log(\ _B) = 0$. If the calibration process is correctly performed, this is equivalent to choosing the minimum-cost threshold for each target language detection sub-system. Thus, after the log-LR transformation, both objective functions to optimize, namely C_{llrAvg} and C_{avg} as defined by NIST [26] tend to be as best as possible. However, a per-language one-vs-all calibration approach as this one will be slightly sub-optimal due to the fact that is does not take into account that this is actually a multiclass problem [27].

V. Databases, protocol and performance metric

LRE'09 evaluation included, for the first time, data coming from two very different audio sources. Besides CTS, used in past evaluations, telephone speech belonging to broadcast news was used for both train and test purposes. Broadcast data was obtained via an automatic acquisition system from "Voice of America" news (VOA) where telephone and non-telephone speech is mixed. Up to 2 terabytes of speech, automatically labeled in language and type, were distributed to participants. Further, around 80 audited segments for each target language (of aproximately 30 seconds duration each) was provided too for development purposes.

Both closed and open-set modes were defined as tasks in this evaluation each one tested with duration segments of 3, 10 and 30 seconds. We refer to closed-set as the task when only target languages are included in the test trials set, and to open-set when other non-target languages (unknown to participants) are also included. In this evaluation, 23 target languages were involved in closed-set as it was shown in Table I and 40 in open-set. More detailed information can be found in the LRE'09 evaluation plan [26].

In order to face this new challenge, where database mismatch play and important role [28], an ATVS development dataset was set up, ATVS-Dev09 onwards. This dataset was built to reproduce in the most accurately possible way, blind evaluation conditions by using different sets of CTS and VOA data provided by NIST. ATVS-Dev09 covered all target evaluation languages and test evaluation duration segments (3, 10 and 30 seconds). Table I shows the 23 evaluation target languages along with ATVS available data type per language. Specifically, the CTS training material (ATVS-DevTrain09) consisted of the "Callfriend" database, the full-conversations of LRE'05 and development data of LRE'07. For Russian data we used also "RuSTeN"[3]. Telephone broadcast data was obtained from speech segments (minimum length 30s.) extracted from VOA long files using telephone labels provided by NIST. The test material (ATVS-DevTest09) was obtained from the test part of LRE'07 (for target languages in both LRE'07 and LRE'09), and from manually labeled data from VOA provided by NIST. Finally, about 15,000 segments, balanced in segments of 3, 10 and 30 seconds, while LRE'09 evaluation included about 15,000 segments per duration (45,000 segments) and therefore about 1 million trials since every segment is tested against every target language.

In order to assess performance, two different metrics were

[2]Available at http://niko.brummer.googlepages.com/

[3]LDC 2006S34 ISBN 1-58563-388-7, www.ldc.upenn.edu

Language	Abbreviation	Data Type (VOA/CTS)
Amharic	*amha*	*VOA/−*
Arabic	*arab*	*−/CTS*
Bengali	*beng*	*−/CTS*
Bosnian	*bosn*	*VOA/−*
Chinese (Cantonese)	*cant*	*VOA/−*
Chinese (Mandarin)	*mand*	*VOA/CTS*
Creole	*creo*	*VOA/−*
Croatian	*croa*	*VOA/−*
Dari	*dari*	*VOA/−*
English (Indian)	*inen*	*−/−*
English (American)	*usen*	*VOA/CTS*
Farsi	*fars*	*VOA/CTS*
French	*fren*	*VOA/−*
Georgian	*geor*	*VOA/−*
German	*germ*	*−/CTS*
Hausa	*haus*	*VOA/−*
Hindi	*hind*	*VOA/CTS*
Japanese	*hind*	*−/CTS*
Korean	*kore*	*VOA/CTS*
Pashto	*pash*	*VOA/−*
Portuguese	*port*	*VOA/−*
Russian	*russ*	*VOA/CTS*
Spanish	*span*	*VOA/CTS*
Tamil	*tami*	*−/CTS*
Thai	*thai*	*−/CTS*
Turkish	*turk*	*VOA/−*
Ukranian	*ukra*	*VOA/−*
Urdu	*urdu*	*VOA/−*
Vietnamese	*viet*	*VOA/CTS*

TABLE I: *Alphabetical list of available languages. In bold, LRE'09 target languages.*

used, both evaluating the capabilities of one-vs.-all language detection. On the one hand, DET curves measure the discrimination capabilities of the system. On the other hand, C_{avg} which is a measure of the cost of taking bad decisions, and therefore it considers not only discrimination, but also the ability of setting optimal thresholds (i. e., calibration). In this work, while DET and C_{avg} results are shown, all our development process was based on C_{avg}, showing now also DET's just to visually observe the discrimination ability of the systems.

VI. Submitted Systems and Notes On Implementation Details

Different combinations of systems presented in Sections II and III were submitted leading to a total of four different systems built under different criteria:

- ***ATVS4*** is a phonotactic-only submission, fusion of the 10 PhoneSVM systems in use (seven from ATVS plus three from BUT)
- ***ATVS3*** is a fast and reliable acoustic-only submission with just the FA-GMM-LS system, designed to optimize the computational time but with a high level of recognition performance.
- ***ATVS2*** consisted of a fusion of all our acoustic (FA-GMM and SVM-SV) and phonotactic (PhoneSVM) systems, as shown in figure 3.
- ***ATVS1*** (primary) is a fusion of ATVS2 with primary system from other participant (TNO), where the latter consisted of a fusion of six acoustic systems: three GMM-SVM and three FA-GMM linear scoring as in [4].

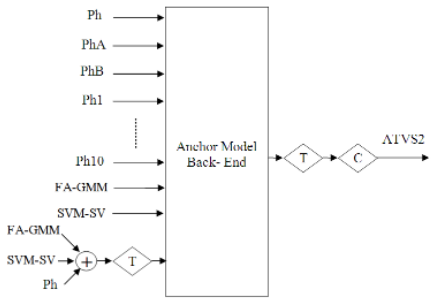


Fig. 3: *Fusion scheme for ATVS2 submitted system.*

A design decision was to generate language models for every target language in both VOA and CTS data where possible depending on data availability, using as well available data on other non-target languages. In that sense 14 CTS and 22 VOA front-end models were trained for every system (VOA IndianEnglish was trained only in the back-end due to data scarcity) as shown in table I. This was done with the goal of later fusing information provided for each model type. Figure 3 shows the fusion scheme for all our systems (ATVS2), remaining fusion systems following similar schemes.

Implementation details for each type of system as well as fusion and calibration notes are shown in the rest of this section.

A. Spectral Systems

A parameterization consisting of 7 MFCCC with CMN-Rasta-Warping [29] concatenated to 7-1-3-7 SDC-MFCCs was used [30] for spectral systems.

According to the data type, two UBMs namely UBM_{CTS} and UBM_{VOA} with 1024 gaussians were trained. Data from CallFriend, LRE'05 and train part of LRE'07 was used for training UBM_{CTS}, while the training of UBM_{VOA} was composed by VOA development data provided by NIST. Distribution per hours of this training is as follows. A total of 38.5 hours was used in UBM_{CTS} training, incluiding about 2.75 hours per 14 available languages. For UBM_{VOA} a total number of 31.2 hours balanced on 1.42 hour per 22 languages was used (IndianEnglish was not included due to data scarcity for this language).

Further, two different FA-GMM-LS systems were developed by using above UBMs. Two session variability subspaces matrices were trained from CTS and VOA data respectively, U_{CTS} and U_{VOA}. We found this approach to outperform the approach where mixed data (CTS,VOA) is processed to train a unique session variability subspace. In this work, session variability subspaces were trained via EM algorithm after a PCA initialization based on [31][7] and only top-50 eigenchannels were taken into account turns out in a $CDx50$ dimension matrix. In order to train the session variability subspaces, a large amount of data was used. U_{CTS} was trained with a total number of 350 hours by using 600 segments of about 150 seconds per the 14 languages available; while U_{VOA} was trained with 550 hours, using 600 segments of

about 150 seconds as well but of the 22 languages available. Data distribution for training UBMs and session variability subspaces is summarized in Table II.

Compensated statistics via Factor Analysis by using U_{CTS} and U_{VOA} as described in II-A3 were also used on the SVM-SV system.

B. High Level Systems

The phonotactic ATVS system is a fusion of 10 different Phone-SVM subsystems (Ph1 to Ph10) as described in Section III. Ph1 to Ph7 use phonetic tokenizers developed by ATVS and Ph8 to Ph10 use phonetic tokenizers trained with Hungarian, Czech and Russian data respectively [4]. The ATVS phonetic tokenizers are based on Hidden Markov Models (HMMs), trained with HTK [32] and later transformed to be used by the SPHINX [33] speech recognition engine for faster recognition. The phonetic HMMs are three-state left-to-right models with no skips, and the output pdf of each state is modeled as a weighted mixture of 20 Gaussians. The acoustic processing is based on 13 Mel Frequency Cepstral Coefficients (MFCCs) (including $C0$) and velocities and accelerations for a total of 39 components, computing a feature vector each 10 ms and performing Cepstral Mean Normalization (CMN). The languages of the phonetic decoders from Ph1 to Ph6 and the corresponding corpora used for training are English (with the corpus with ELDA catalogue number S0011), German (S0051), French (S0185), Arabic (S0183 + S0184), Basque (S0152) and Russian (S0099)[5]. Ph7 uses a phonetic decoder in Spanish trained on Albayzin spanish speech database [34] downsampled to 8 kHz, which contains about 4 hours of high-quality phonetically labelled speech. Once the speech segment has been transformed into a sequence of recognized phonetic tokens (with any of the phonetic decoders), this sequence is used to estimate count-based 1-grams, 2-grams and 3-grams, pruned with a probability threshold, resulting in about 40,000 n-grams. These are rearranged as a feature vector, which is taken as the input of an SVM that classifies the test segment as corresponding (or not) to one language. PhoneSVMs are combined in different ways to obtain different front-end systems. Each PhX system consists of 22 VOA and 14 CTS models trained separately. Channel dependent t-norm is the last stage of those phonotactic front-ends.

C. Fusion and calibration

Input vectors to our fusion systems anchor model based back-end had dimension 216 (36 ATVS models - 14CTS+22VOA- x 6 component systems) while primary was 438 adding scores output of other site. Back-end t-norm was design as channel-independent (VOA+CTS), while calibration was duration-dependent. Anchor model training was 90/10 bootstrapped while calibration training was bootstrapped with 80/20 using available training data. A channel independent T-Norm (models from VOA and CTS) stage was applied for scoring normalization.

LRE'09 considered three different nominal durations for the test segments: 3, 10 and 30 seconds of speech. The same individual subsystems were used to perform language recognition tests for the different durations. However, calibration has been trained specifically for the estimated different durations and an automatic voice activity detector has been used to classify test segments. As the calibration was applied after the back-end, a single score for each test segment was used, and scores from all the speech types (VOA, CTS) were pooled for training. Thus, all the available scores for each duration from each target language were used to train logistic regression, and the linear transformation obtained was used to calibrate the scores from testing data.

VII. Development and evaluation Results

The performance of ATVS submitted systems is summarized in Figure 4 for development (ATVSDev09) and evaluation (LRE'09) tests. Here, the discrimination per each system (ATVS1-4) and test segment duration (3, 10 and 30 seconds) is showed in a pooled DET curve. Several global observations can be inmediately extracted. Firstly, the good behaviour of the anchor models fusion scheme introduced is justified as being ATVS1 (fusion of systems) the system with lower error rates. The effect of test segment duration in system performance is also highlighted and it affects in a similar manner to both, acoustic and high level systems. Further, a slight degradation in the evaluation results with respect to development ones is showed. This degradation performance, common to all participants, is usually due to the database mismatch among the development and testing databases, and is a common effect in LRE's. Table III summarizes this information in terms of *meanCavg* (mean of Cavg per language) per system, evaluation dataset and test segment durations. It is also worth pointing out that acoustic systems outperform phonotactic ones except for short durations, and this with a much smaller computational complexity, but fusion of both kind of systems improve results, which encourages the use of multilevel approaches for language recognition.

In more detail, Figure 5 compares systems performance per target language. Again, results are presented on both, development and evaluation, but only for 30s test segment duration. Analysis shows the varying degrees of recognition difficulty among the different target languages (or better said, among the data available from those target languages). In the same way, Figure 6 presents in detail the effect of test segment duration per language for our primary system (ATVS1).

The need of proper session variability compensation is showed in Figure 7 where both spectral systems, FA-GMM-LS and SVM-SV are assesed with and without compensation via factor analysis on ATVSDev09. Results shows that channel compensation via FA is crucial in GMM modelling performance, getting an improvement of about 82% in $meanC_{avg}$ terms. Also, system SVM-SV take advantage of this compensation but to a lesser extent (4%). This effect appears

[4]These have been developed and made available for research purposes by the Speech Processing Group at Faculty of Information Technology, Brno University of Technology.

[5]www.elda.org.

Prior model	Databases	#Languages	#Hours/language	Total
UBM_{CTS}	$CallFriend, LRE05, TrainLRE07$	14	2.75	38.5
U_{CTS}	$CallFriend, LRE05, TrainLRE07$	14	25	350
UBM_{VOA}	VOA	22	1.42	31.2
U_{VOA}	VOA	14	25	550

TABLE II: *Distribution of data used for training Universal Background Models and Session Variability Subspaces.*

	ATVS Systems Performance					
	ATVS-Dev09			LRE'09		
	03s	10s	30s	03s	10s	30s
$ATVS1$	16.50	**6.48**	**1.56**	17.97	**7.87**	**3.71**
$ATVS2$	**16.17**	7.25	2.02	**17.92**	8.39	4.26
$ATVS3$	20.37	10.30	3.25	21.93	10.65	5.67
$ATVS4$	18.80	9.41	3.73	20.87	10.81	6.55

TABLE III: *ATVS submitted systems performance (meanCavg x 100) on development and evaluation datasets.*

	ATVS1 on LRE09		
	$03s$	$10s$	$30s$
$closed-set$	17.97	7.87	3.71
$open-set$	18.69	8.80	4.58

TABLE IV: *ATVS1 performance (meanCavg x 100) on LRE'09 closed- and open-set.*

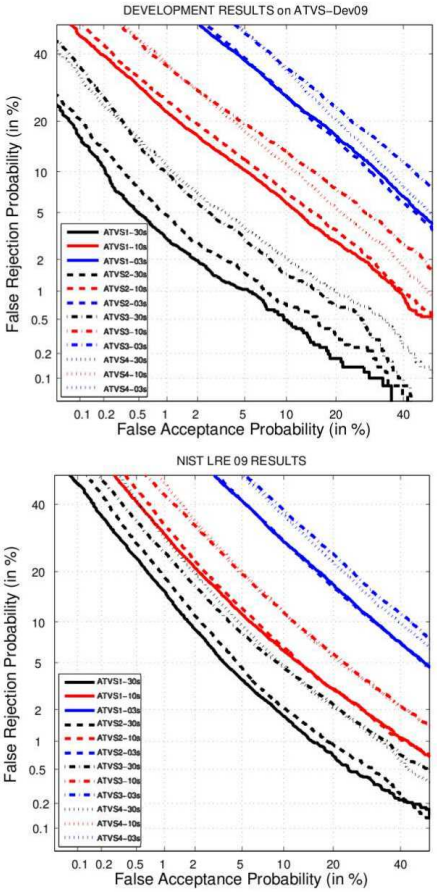


Fig. 4: *Pooled DETs per ATVS submitted systems on development (ATVS-Dev09) and evaluation (LRE'09) per all target test segment durations (3, 10 and 30 seconds)*

due to differences in SVM and GMM modelling. In GMM, target languages models, trained with huge amount of data, are far shifted with respect UBM reference model after even a single MAP adaptation. This mean shifting includes not only information belonging to the language but session variability found in the training database which it is mainly independent of the languages. This leads to models that are growing strongly affected by session variability effects. On the contrary, the SVM exhibits a higher robustness to this problem due to its ability to estimate an hyperplane separating target single utterances models against all non-target ones. However, once session variability compensation is applied, GMM outpeforms SVM-SV system.

Table VII presents the system performance of our primary system on the closed- and open-set where a total of 40 languages were involved (23 target + 17 non-target). Results for the core condition (closed-set, 30s) are comparable to the best systems in the evaluation. It is worth highlighting the excellent performance of the ATVS primary system in the open-set condition, where a second rank position was obtained. Results in that task prove the robustness of anchor models working under unseen languages.

VIII. Conclusion and Future Work

In this article we have described the ATVS-UAM submission to the 2009 NIST Language Recognition Evaluation. This submission was particularly successful since our systems achieved the 2^{nd} position in the open-set condition with speech segments of 30 seconds. The article has discussed and presented the state-of-the-art technologies used in our systems, with emphasis on the two main research innovations introduced. Firstly, anchor models based fusion has been proposed and has proven to be an excellent scheme for fusion of a set of different subsystems. Secondly, session variability compensation has been applied on statistics domain and has shown to outperform the SVM-SV system, thus avoiding the need for a frame by frame compensation and allowing statistics extracted from the linearized FA-GMM system to be reused. Besides these innovations, the LRE'09 task included several new research challenges with respect to former evaluations, as huge amount of data to process and a larger number of target languages (23). A special mention deserves the broad session variability due to the use of telephone data from two different sources, broadcast news (extracted from Voice of America news -VOA-) and conversational telephone speech (CTS). ATVS acoustic and high level systems were built taking into account all these factors and achieved good performance

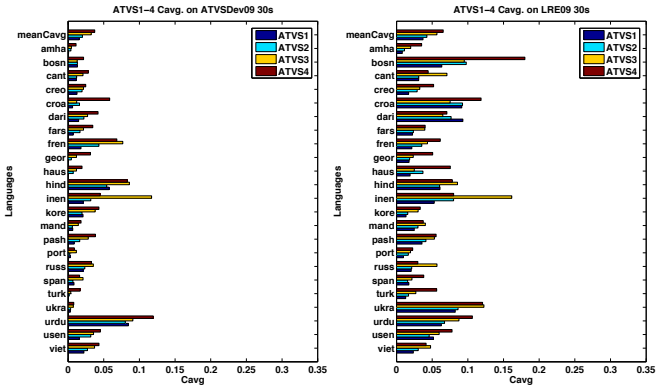


Fig. 5: *Comparision of ATVS submitted systems on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets for 30 seconds test duration segments.*

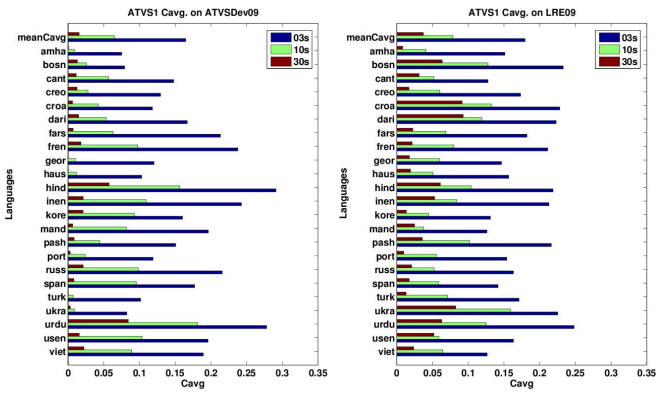


Fig. 6: *ATVS primary system performance on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets (3, 10 and 30s).*

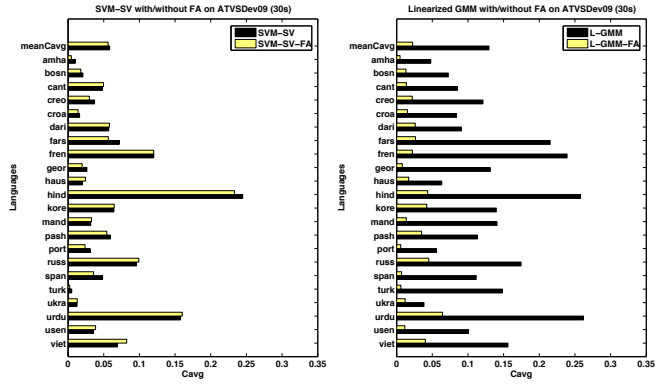


Fig. 7: *Effect of session variability compensation on SVM-SV and FA-GMM-LS systems. Results on ATVS-Dev09 using VOA models and* U_{VOA}.

in the task with remarkable results in all submitted tasks. To achieve this goal, the use of a powerful session variability compensation scheme via Factor Analysis have demonstrated to be crucial for acoustic systems performance, obtaining significant improvements in both the SVM-SV and the FA-GMM-LS models submitted. Future work includes several lines such as to explore new accurate ways to better extract and combine complementary information from different systems; to build systems more independent to the effects of test duration and to explore new techniques for fast adaptation to new channel conditions in session variability compensation when a limited set of unseen background data is available.

ACKNOWLEDGMENT

This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. Javier Gonzalez-Dominguez also thanks Spanish Ministry of Education for supporting his doctoral research under project TEC2006-13141-C03-03. Special thanks are given to Dr. David Van Leeuwen from TNO Human Factors (Utrech, The Netherlands) for his strong collaboration, valuable discussions and ideas. Also, authors thank to Dr. Patrick Lucey for his final support on (non-target) Australian English review of the manuscript.

REFERENCES

[1] K. Atkinson, "Language Identification from Nonsegmental Cues," *The Journal of the Acoustical Society of America*, vol. 44:378A, 1968.

[2] National Institute of Standards and Technology, "NIST LRE website," http://www.nist.gov/speech/tests/lang, (accessed 04 July 2008).

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[4] N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and O. Glembek, "Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics," in *Proc. of Interspeech*, 2009.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 637–640.

[6] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[7] R. Vogt and S. Sridharan, "Explicit Modeling of Session Variability for Speaker Verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[8] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support Vector Machines for Speaker and Language Recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[9] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, "High-level Speaker Verification with Support Vector Machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 73–76.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[11] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[12] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. 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.

[13] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of Scoring Methods Used in Speaker Recognition with Joint Factor Analysis," in *ICASSP '09: Proc. of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4057–4060.

[14] P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, "Speaker and Session Variability in GMM-based Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[16] J. Gauvain and C. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian mixture Observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[17] V. Wan and S. Renals, "Speaker Verification Using Sequence Discriminant Support Vector Machines," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 203–210, March 2005.

[18] W. M. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines Using a GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[19] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," in *ICASSP*, vol. I, 2005, pp. 629–632.

[20] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification," in *Proc. of Interspeech*, 2007, pp. 1242–1245.

[21] W. M. Campbell, D. Sturim, P. Torres-Carrasquillo, and D. Reynolds, "A Comparision of Subspace Feature-Domain Methods for Language Recognition," in *Proc. of Interspeech 2008*, September 2008.

[22] M. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[23] I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez, and D. T. Toledano, "Anchor-model Fusion for Language Recognition," in *Proc. of Interspeech 2008*, September 2008.

[24] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic Anchor Models for Speaker Verification," in *Proc. of Interspeech*, vol. 1, 2005, pp. 211–214.

[25] E. Noorl and H. Aronowitz, "Efficient Language Identification Using Anchor Models and Support Vector Machines," in *Speaker Odyssey*, 2006, pp. 1–6.

[26] "The 2009 NIST Language Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf (accessed 20 July), 2009.

[27] N. Brümmer and D. van Leeuwen, "On Calibration of Language Recognition Scores," in *Proc. of Odyssey*, San Juan, Puerto Rico, 2006.

[28] D. Ramos, J. Gonzalez-Rodriguez, and J. Gonzalez-Dominguez, J. Lucena, "Adressing Database Mismatch in Forensic Speaker Recognition with Ahumada III: a Public Real-Casework Database in Spanish," in *Proc. of Interspeech*, 2008.

[29] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp. 213–218.

[30] A. Rosenberg, C. Lee, and F. Soong, "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," in *ICSLP*, vol. 1, 2002, pp. 89–92.

[31] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling With Sparse Training Data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[32] "Hidden Markov Model Toolkit (HTK)," available on http://htk.eng.cam.ac.uk/.

[33] "Carnegie Mellon University SPHINX speech recognizer," available on http://sourceforge.net/projects/cmusphinx/.

[34] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mario, and C. Nadeu, "ALBAYZN Speech Database: Design of the Phonetic Corpus," in *European Conference on Speech Communication and Technology, Eurospeech*, vol. 1, 1993, pp. 175–178.

Javier Gonzalez-Dominguez received his M.S. degree in Computer Science in 2005 from Universidad Autonoma de Madrid, Spain. In 2007 he obtained the postgraduate Master in Computer Science and Telecommunication Engineering from U.A.M. He has formed part of other international research groups and has participated in the development of the ATVS speaker and language recognition systems for several NIST speaker and language recognition evaluations, he has also participated on several related projects since 2005.

Ignacio Lopez-Moreno received his M.S. degree in Telecommunication Engineering in 2009 from Universidad Politecnica de Madrid (UPM). Currently he is a PhD student with the Biometric Recognition Group - ATVS -, where he is working as an assistant researcher since 2004. He has participated in several national projects and technology evaluations, and NIST speaker and language recognition since 2005. He has been recipient of several awards and distinctions, such as the IBM Research Best Student Paper in 2009.

Javier Franco-Pedroso Bachelor degree in Telecommunication Engineering (image and speech intensification) in 2005 from Universidad Politecnica de Madrid (UPM). Since september 2004 he is as an assistant researcher for the Biometric Recognition Group - ATVS - at Universidad Autonoma de Madrid. Currently he is an Electrical Engineering master degree student at UPM. His research interests include speaker verification, speaker tracking, pattern recognition and speech processing.

Daniel Ramos received his M.S in Telecommunication Engineering in 2001 from Universidad Politecnica de Madrid, Spain; and Ph.D. in Telecommunication Engineering in 2007 from Universidad Autonoma de Madrid, Spain. He is, since 2003, with ATVS - Biometric Recognition Group -, and since 2006 works as an Assistant Professor at Universidad Autonoma de Madrid. Dr. Ramos has participated in the development of the ATVS speaker and language recognition systems since 2004. He has been part of several organizing and scientific committees in the field, and has been recipient of several awards and distinctions, such as the IBM Research Best Student Paper Award at Odyssey 2006.

Dorote Torre Toledano M.S. Telecommunication Engineering degree from the Universidad Politecnica de Madrid, Spain, in 1997, obtaining the best academic records of his class. Ph.D. degree in telecommunication engineering from the same university, receiving a Ph.D. Dissertation Award from the Spanish Association of Telecommunication Engineers. He is with ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid (Spain) where he is currently Associate Professor. He has served as a member of the scientific committee of several international conferences as well as a reviewer for several journals in the field.

Joaquin Gonzalez-Rodriguez , received the M.S. degree in 1994; and the Ph.D. degree "cum laude" in 1999, both in electrical engineering, from Univ. Politecnica de Madrid (UPM), Spain. Dr. Gonzalez-Rodriguez is co-director of the Biometric Recognition Group - ATVS -. After 15 years of research and lecturing at UPM, he is since May 2006 an Associate Professor at the Computer Science Department at Univ. Autonoma de Madrid, Spain. He has led ATVS participations in NIST Speaker and Language Recognition Evaluations since 2001. He is a member of ISCA and the Signal Processing Society of IEEE, and is also a member of the Program Committee of the ISCA Odyssey conferences on Speaker and Language Recognition.

3.5. Fine-grained automatic speaker recognition using cepstral trajectories in phone units

Título: "Fine-grained automatic speaker recognition using cepstral trajectories in phone units"
Autores: Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez, Javier Gonzalez-Dominguez and Daniel Ramos
Libro: Quantitative approaches to problems in linguistics
ISBN: 9783862883844 (Hardbound)
Editor: LINCOM Studies in Phonetics 08, 2012

Fine-grained automatic speaker recognition using cepstral-trajectories in phone units*

Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez, Javier Gonzalez-Dominguez and Daniel Ramos
Universidad Autonoma de Madrid

In this paper, the contributions to speaker identity from different phone units are explored through the analysis of the temporal trajectories of their Mel-Frequency Cepstral Coefficients (MFCC). Inspired by successful work in forensic speaker identification, we extend the approach based on temporal contours of formant frequencies in linguistic units to design a fully automatic system, bringing together both forensic and automatic speaker recognition worlds. The combination of MFCC feature extraction and variable-length unit-dependent trajectories coding provides a powerful tool to extract individualizing information. At a fine-grained level, we provide a calibrated likelihood ratio per linguistic unit under analysis (extremely useful in applications such as forensics), and at a coarse-grained level, we combine the individual contributions of the different units to obtain a single calibrated likelihood ratio per trial. This approach has been tested with datasets and protocols used in the 2006 Speaker Recognition Evaluation (SRE) carried out by the US National Institute of Standards and Technology (NIST), consisting of 9,720 trials from 219 male speakers for the 1side-1side English-only task, and development data being extracted from 367 male speakers from 1,808 conversations from NIST SRE 2004 and 2005 datasets.

1. Introduction

Automatic speaker recognition has focused in the last decade on two concurrent problems: the compensation of session variability effects, mainly through high-dimensional supervectors and latent variable analysis (Dehak et al. 2011; Kenny et al. 2008; Kenny 2010), and the production of an application-independent calibrated likelihood ratio per speaker recognition trial (Brummer & du Preez 2006), able to elicit useful speaker identity information to the final user with any given application prior. The results are highly efficient text-independent systems in controlled conditions, as NIST SRE evaluations, where lots of data from hundreds of speakers in similar conditions are available. Thus, all the speech available in every trial is used to produce detection performances difficult to imagine a decade ago.

However, in the presence of strong mismatch (as e.g. in forensic conditions, where acoustic and noise mismatch, apart from highly different emotional contexts, speaker roles or

*Supported by MEC grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10-UAM/TIC-5792 and Catedra UAM-Telefonica. Thanks to ICSI (Berkeley, CA) for hosting the preliminary part of this work. Thanks to SRI for providing Decipher phone, word and syllable labels for SRE04 and SRE06.

health/intoxication states can be present between the control and questioned speech), those acoustic/spectral systems could be unusable as all our knowledge about the two speech samples is deposited into a single likelihood ratio, obtained from all the available speech in the utterance, that could be strongly miscalibrated (being then highly misleading) as the system has been developed under severe database mismatch between training and testing data. Moreover, it is difficult (or even impossible) to collect enough data to develop a system robust to every combination of mismatched factors present in actual case data, an important problem in real applications.

A usual procedure in forensic laboratories is that a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable speech between both samples, segments being from seconds long to just some short phonetic events in given articulatory contexts. The number and types of comparable units for analysis is always a case-dependent subject, and therefore flexible strategies for analysis and combination are needed.

The proposed approach proposes an improvement to this application framework, providing informative calibrated likelihood ratios for every linguistic unit under analysis. Moreover, the combination of the different units yields good discrimination capabilities allowing one to obtain speaker detection performance levels similar to equivalent acoustic/spectral systems when enough usable units are available.

The remainder of the paper is organized as follows. In Sections 2 and 3 we presentour proposed front-end for feature extraction over phone units and the system in use. Section 4 describes the databases and the experimental protocol used for testing the system. Section 5 shows results for the different phone units individually and for several combination methods, to finally conclude in Section 6 summarizing the main contributions and future extensions of this work.

2. Uniform-length feature extraction from variable-length speech segments

Many attempts have been made to incorporate the temporal dynamics of speech into features, from the simplest use of the velocity (delta) and acceleration (delta-delta) derivative coefficients to modulation spectrograms, frequency modulation features or even TDCT (temporal DCT) features (see Kinnunen & Li, 2010 for a review). However, to the best of our knowledge none of the previous approaches, with the exception of SNERFs (Ferrer 2009), and Shriberg et al. (2005) for prosodic information, take advantage of the linguistic knowledge provided by an automatic speech recognizer (ASR) to extract non-uniform-length sequences of spectral vectors to be converted into constant-size feature vectors characterizing the spectro-temporal information in a given linguistic unit. Similar approaches, although based on F-patterns, have used linguistic information to perform forensic speaker recognition (Gonzalez-Rodriguez et al. 2007), motivating this work. In our proposed front-end, we obtain a constant-size feature vector from non-uniform-length MFCC features' sequence within a phone unit.

2.1 ASR region conditioning

In order to extract the phone units from the speech signal, the phonetic transcription labels produced by SRI's Decipher conversational telephone speech recognition system (Kajarekar,

et al., 2009) were used first. For this system, which was trained on English data, the Word Error Rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively. These labels define both phonetic content and time interval of speech regions containing phone units to be segmented. For this work, 39 phone units from an English lexicon were used, represented by the Arpabet phonetic transcription code (Wikipedia). The equivalence between the Arpabet phonetic symbols and the International Phonetic Alphabet (IPA) is shown in Table 1. In addition, two filled pauses commonly used (coded as 'PUH' and 'PUM') were also added to this set of phones for speaker recognition purposes.

2.2 Cepstral-trajectories parameterization

By means of SRI's Decipher phone labels, trajectories (i.e., the temporal evolution of each MFCC vector dimension) of 19 static MFCC are extracted, yielding a MFCC matrix of 19 coefficients x #frames/phone for each phone unit. This variable-length segment is then duration equalized to a number of frames equivalent to 250 ms by means of an interpolation/decimation process. Finally, those trajectories are coded by means of a fifth order discrete cosine transform (DCT), yielding our final 19 x 5 fixed-dimension feature vector for each phone unit.

3. System description

3.1 Phone-dependent acoustic systems

Our proposed system is based on the GMM-UBM framework (Reynoldset al. 2000), widely used not only in the automatic speaker recognition field but also known by the forensic community (Rose & Winter, 2010), using duration-equalized DCT-coded MFCC trajectories per phone unit as feature vectors. One single 1024-mixtures gender-dependent UBM is trained using all existing phone units per utterance for each individual from a background population (development dataset). Then, for each target speaker in the evaluation dataset, a speaker model is generated by means of *maximum a posteriori* (MAP) adaptation, using all the phone units available. Finally, the scoring process is performed following a phone-dependent scheme, using only feature vectors belonging to the phone unit under analysis. This procedure yields N scores per trial (N=#phones) that can be either used as individual speaker recognition systems or, by contrast, combined in a single fused system. On the one hand, individual phone-unit systems allow us to report useful speaker verification LR's for very short speech samples where the usual state-of-the-art automatic speaker recognition systems are not directly applicable (as is the case for forensic applications). On the other hand, when more data is available, individual phone units can be combined to achieve better discriminative capabilities.

3.2 Fusion schemes

In addition to obtaining test results for each phone unit, these individual systems were combined. First, scores-sets belonging to each individual system were calibrated by means of linear logistic regression. Then, the sum fusion rule was applied for simplicity (there are 41 systems and a lot of combinations were analyzed, so a technique that involves little

computational load was needed), but other techniques such as logistic regression fusion could be applied to obtain better combination results.

Another issue is which phones should be selected for fusing. Two strategies have been used in this work. The first is to select the n-best performing phones by setting a threshold for the Equal Error Rate (EER) of the individual systems to be fused, leaving out the poorer performing phones. However, this procedure does not guarantee that the best fused system will be achieved because some phones with lower performance could contribute to the fused system if its LR's have a sufficiently small correlation coefficient with those produced by the other phones to be fused. On the other hand, testing all of the possible combinations would be a very complex task, so we used a phone selection algorithm (similar to that used in Castroet al., 2009) based on the following steps:

- Take the best performing phone in terms of EER as the initial phones set.
- Take the next best performing phone and fuse with the previous set. If the fusion improves the performance of the previous set, this phone is added to the phones set, otherwise rejected.
- The previous step is repeated for all the units in increasing EER order.

This procedure allows us to find complementarities between phones that otherwise would not have been revealed, but avoiding the complex task of testing each possible combination.

4. Datasets and experimental setup

4.1 NIST SRE databases and protocols

The US National Institute of Standards and Technology (NIST) has been conducting Speaker Recognition Evaluations (SRE) from 1997 in order to measure the state-of-the-art and to find the most promising algorithmic approaches in text-independent speaker recognition. These evaluations define datasets and protocols to measure system performance in an objective way, so that the results obtained by systems based on very different technologies can be compared in a common framework.

Briefly, for each SRE, NIST provides an evaluation dataset consisting of two subsets: a training dataset containing excerpts of target speakers to be modeled, and a test dataset containing test segments from unknown individuals to be compared with target speaker models. Several conditions in duration (10 sec., 30 sec., 2.5 min., etc.) and audio types (several telephone and microphone channels) for training and test segments are established, so that different combinations between them define different tasks to be faced. For this work, only the 1side-1side task has been evaluated, consisting of 5 minutes telephonic conversations (of approximately 2.5 minutes per conversation side) for both training and testing stages.

4.2 Development and evaluation datasets

NIST SRE datasets and protocols have been used to develop and test our proposed system, in particular those of years 2004, 2005 and 2006. As region conditioning for phone units definition and extraction rely on SRI's Decipher ASR system (trained on English data), English-only subsets of the NIST SRE datasets have been used. SRE 2004 and 2005 datasets were used as the background dataset for UBM training, consisting of 367 male speakers from 1,808 conversations (only male speakers were used for this work). The 1side-1side task from

the SRE 2006, involving only English male speakers, was used for testing purposes. This dataset and evaluation protocol comprises both native and nonnative speakers across 9,720 same-sex different-telephone-number trials from 298 male speakers. The SRE 2005 evaluation set was also used to obtain scores in order to train the calibration rule (linear logistic regression).

The performance evaluation metrics used are the EER and the Detection Cost Function (DCF) as defined in the NIST SRE 2006 evaluation plan (National Institute of Standards and Technology 2006). C_{llr} and $minC_{llr}$ (Brummer & du Preez 2006) (and its difference, calibration loss) are also used to evaluate the quality of the different detectors after the calibration process.

5. Results

5.1 Individual phone unit performance

Table 1 shows the individual performance of phone units for the NIST SRE 2006 English-only male 1side-1side task, while figure 1 shows their Detection Error Tradeoff (DET) curves (only the 11 best perfoming phones are highlighted). It can be seen that, although most of the phones have high EER and minDCF values, almost all of them are well calibrated (low difference between C_{llr} and $minC_{llr}$), as we can also see in the tippet plot in figure 2 for the best performing phone unit ('N'). This enables informative calibrated likelihood ratios to be obtained from very short speech samples (as low as some phone units). On the one hand, it should be noted that, for each individual system, we are using a very small amount of testing data compared with our reference system (GMM-UBM based on 19 MFCC + deltas feature vectors, whose performance is EER=10.26% and minDCF=0.0457) which uses the entire test file (around 2.5 minutes of speech). On the other hand, the phone labeling and time intervals annotation rely on a fully automatic speech recognition system, so errors made by this system affect the phone-based speaker recognition systems performance. However, there are a lot of phone units that can be combined to achieve better discrimination capabilities of the overall system, as it is shown in further discussion.

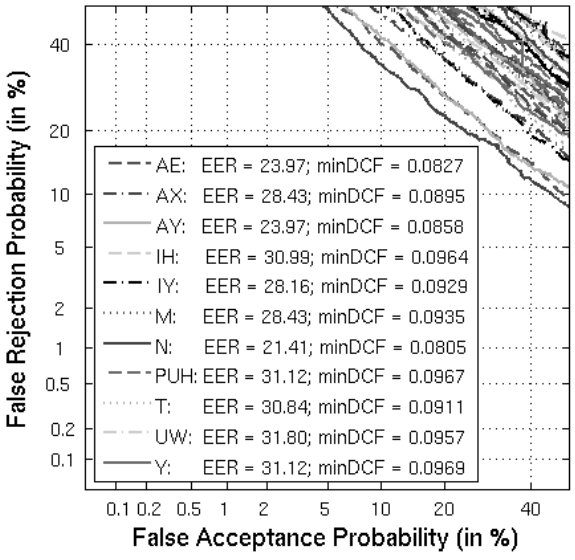


Figure 1. DET curves for individual phone unit performance in the NIST SRE 2006 English-only male 1side-1side task. EER and minDCF are shown for the eleven best performing phones.

Arpabet	IPA	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AO	ɔ	39.90	0.1000	0.9488	0.9316
AA	ɑ	37.33	0.0993	0.9284	0.9137
IY	i	28.16	0.0929	0.8072	0.7810
UW	u	31.80	0.0957	0.8681	0.8468
EH	ɛ	33.56	0.0975	0.8919	0.8679
IH	ɪ	30.99	0.0964	0.8500	0.8252
UH	ʊ	42.49	0.1000	0.9851	0.9654
AH	ʌ/ə	31.94	0.0972	0.8825	0.8468
AX	ə	28.43	0.0895	0.8161	0.7792
AE	æ	23.97	0.0827	0.7230	0.6988
EY	eɪ	32.48	0.0935	0.8734	0.8480
AY	aɪ	23.97	0.0858	0.7347	0.7044
OW	oʊ	33.83	0.0967	0.8829	0.8626
AW	aʊ	40.17	0.1000	0.9549	0.9409
ER	ɝ/ɚ	39.24	0.1000	0.9484	0.9298
P	p	42.59	0.1000	0.9763	0.9683
B	b	38.55	0.0995	0.9453	0.9311
T	t	30.84	0.0912	0.8291	0.8035
D	d	36.03	0.0986	0.9056	0.8885
K	k	33.96	0.0989	0.8932	0.8721
G	g	42.46	0.1000	0.9790	0.9654
CH	tʃ	43.53	0.1000	0.9866	0.9708
JH	dʒ	42.52	0.0997	0.9809	0.9695
F	f	43.00	0.1000	0.9796	0.9701
V	v	42.07	0.1000	0.9678	0.9589
TH	θ	42.19	0.0998	0.9768	0.9662
DH	ð	35.58	0.0973	0.9061	0.8847
S	s	35.18	0.0966	0.8861	0.8708
Z	z	37.61	0.0997	0.9368	0.9234
SH	ʃ	43.41	0.0998	0.9834	0.9736
HH	h	44.35	0.0996	0.9837	0.9673
M	m	28.42	0.0936	0.8198	0.7930
N	n	21.41	0.0806	0.6811	0.6546
NG	ŋ	34.82	0.0993	0.9192	0.8927
L	l	34.89	0.0981	0.8981	0.8804
R	r *or* ɹ	34.72	0.0973	0.8986	0.8797
DX	ɾ	40.34	0.0997	0.9686	0.9527
Y	j	31.27	0.0969	0.8598	0.8365
W	w	40.44	0.0997	0.9639	0.9531
PUH		31.12	0.0967	0.8381	0.8185
PUM		38.95	0.0985	0.9229	0.9019

Table 1. EER (%), minDCF, C_{llr} and $minC_{llr}$ for phone units in the NIST SRE 2006 English-only male 1side-1side task.

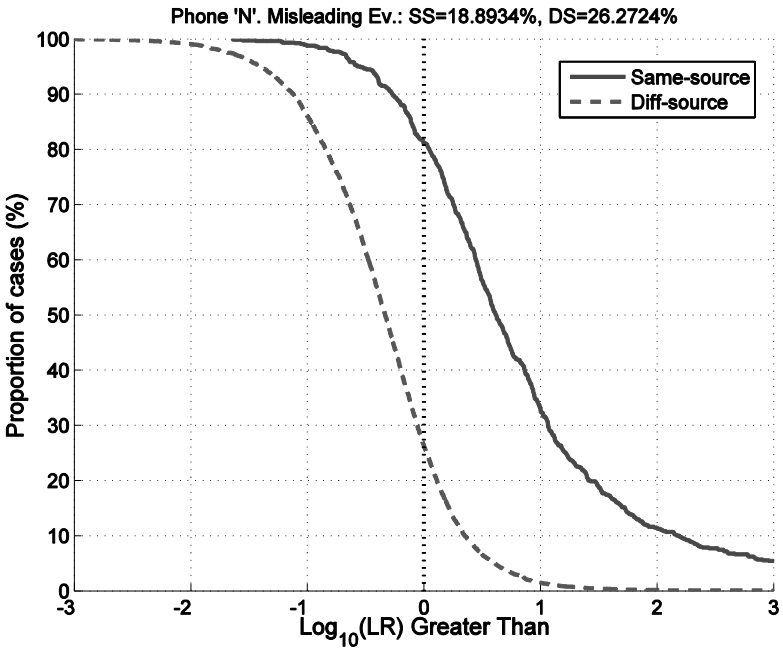


Figure 2. Tippet plot for the best performing phone unit ('N') in the NIST SRE 2006 English-only male 1side-1side task.

5.2 Phone units combination

The first of the two fusion schemes proposed in section 3.2 (setting a threshold for the EER of the phone units to fuse) yields the results shown in the first six entrances of table 2 and figure 3. It can be seen that: i) with few phones fused (3-best fusion), system performance is highly improved with respect to that of the individual phones (30% improvement with respect to the best phone fused, in terms of EER), showing very high complementarity between phones; and ii) this complementarity is confirmed as more phones are added to the fusion, improving system performance till it stabilizes close to that of the reference system; as a side effect, system calibration degrades as more phones are added to the fusion (figures 4 and 5). This effect is due to the use of the sum fusion rule: when individual systems are fused, the errors made by them are added given a higher misleading LR. This could be avoided by using a different fusion technique such as the average rule, so that the fused LR's wouldn't be amplified.

Using the second fusion scheme (the phone selection algorithm), a lower EER can be achieved (last entry in table 2 and figure 3), using a set of 11 phone units. It is worth noting that some of the 13-best performing phones are not included in this set ('AX', 'EY', 'IH', 'T' and 'Y'), while others with lower performance are ('AO', 'NG' and 'UH'). However, calibration metrics are significantly worse than those of the 13-best performing phones fusion.

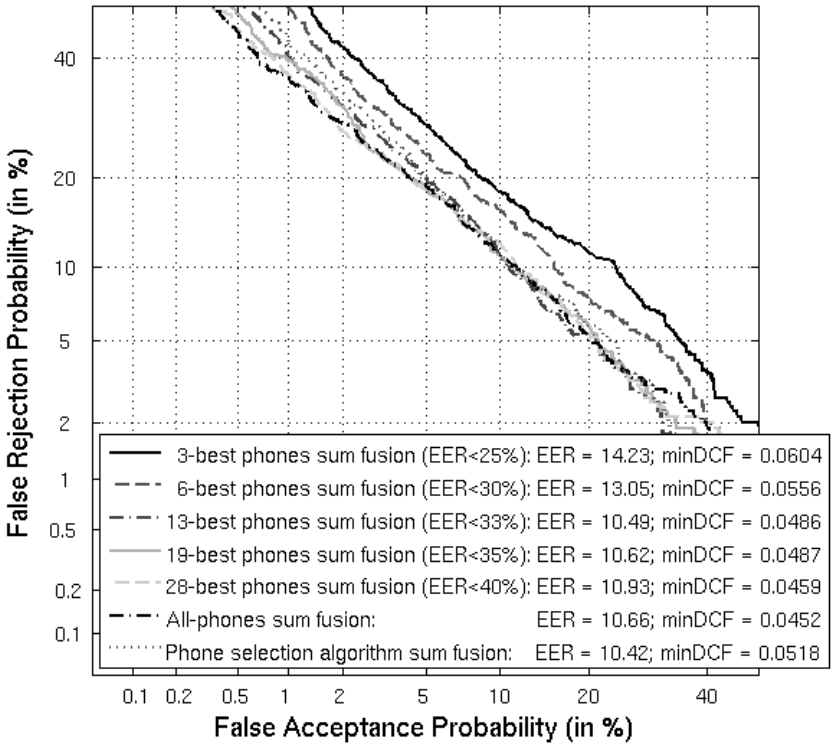


Figure 3. DET curves for different fusion schemes in the NIST SRE 2006 English-only male 1side-1side task.

EER threshold	#phones	EER (%)	minDCF	C_{llr}	$minC_{llr}$
25	3	14.23	0.0604	0.5133	0.4604
30	6	13.05	0.0556	0.5283	0.4065
33	13	10.49	0.0486	0.5943	0.3535
35	19	10.62	0.0487	0.6957	0.3544
40	28	10.93	0.0459	0.7829	0.3532
-	41	10.66	0.0452	0.8593	0.3540
-	11	10.42	0.0518	0.6895	0.3618

Table 2. EER (%), DCF, Cllr and minCllr for sum of phone units in the NIST SRE 2006 English-only male 1side-1side task.

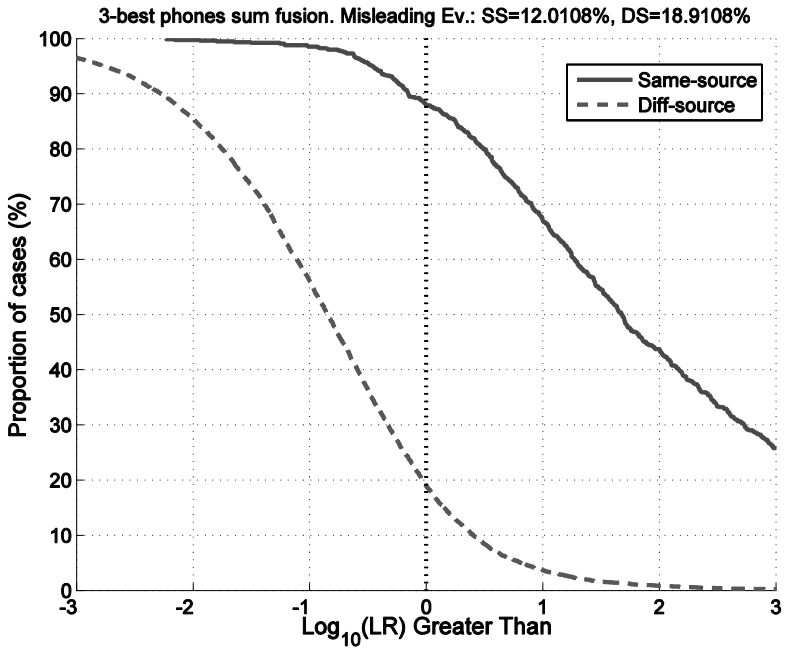


Figure 4. Tippet plot for the sum fusion of 3-best performing phone units ('AE', 'AY' and 'N') in the NIST SRE 2006 English-only male 1side-1side task.

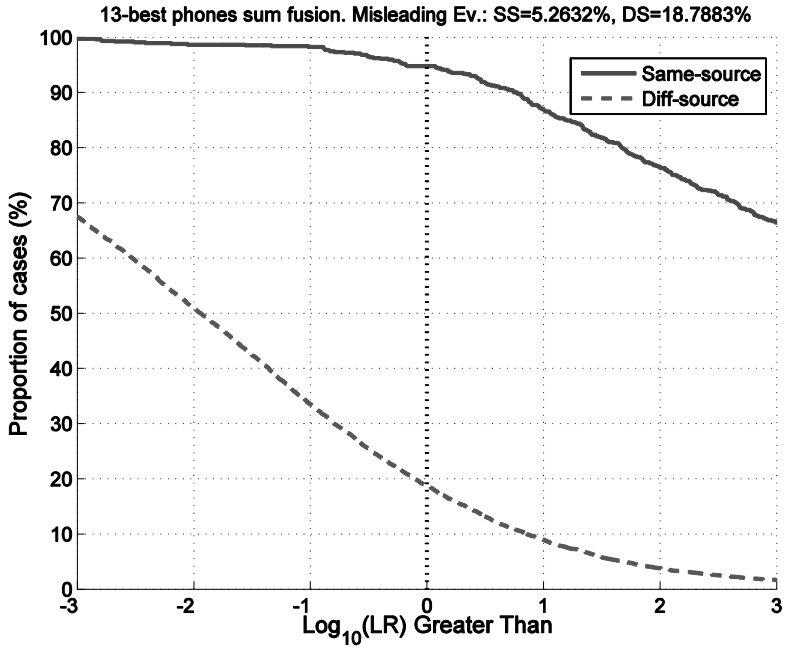


Figure 5. Tippet plot for the sum fusion of 13-best performing phone units ('AE', 'AH', 'AX', 'AY', 'EY', 'IH', 'IY', 'M', 'N', 'PUH', 'T', 'UW' and 'Y') in the NIST SRE 2006 English-only male 1side-1side task.

5.3 MFCC and cepstral-trajectories systems fusion

Finally, to analyze the complementarity between our proposed phone-unit systems (based on cepstral-trajectories) and clasical speaker recognition systems (based on MFCC features), a sum fusion was performed between our reference system and the best performing phones fusion found in experiments reported in the preceding paragraph (the 13-best phones sum fusion). Results are shown in Figure 6, where it can be seen that it is possible to obtain a consistent improvement over the reference system at every system operating point. So, our phone-unit systems are not only useful in cases where usual automatic speaker recognition systems are not directly applicable, but also in cases where they are, improving the performance of the overall system.

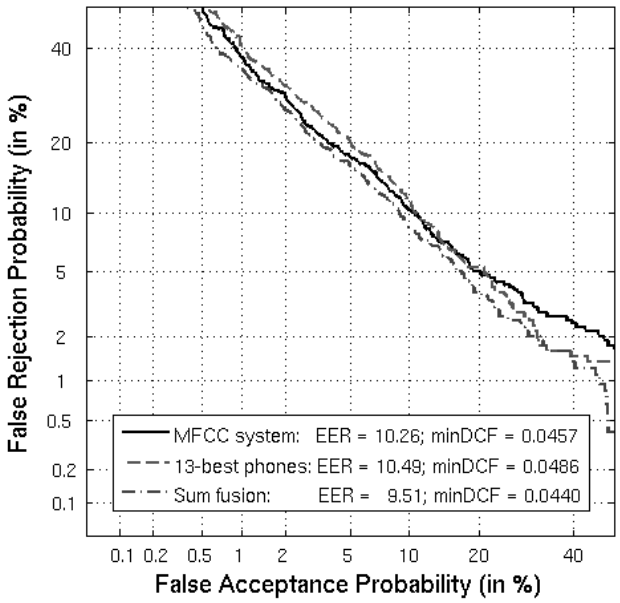


Figure 6. DET curves for MFCC system, cepstral-trajectories system and their sum fusion in the NIST SRE 2006 English-only male 1side-1side task.

6. Summary and conclusions

In this paper we have presented an analysis of the contributions of individual phone units to automatic speaker recognition by means of their cepstral-trajectories, showing that some of them exhibit perfectly acceptable performance and likelihood ratios informative to forensic applications but with the advantage of being a completely automatic system. This way it is possible to deal with uncontrolled scenarios where only some short segments are available to be compared, making it possible to infer a conclusion about the speaker identity in the speech sample. This procedure cannot be done by the usual automatic speaker recognition systems because they use all available speech data as a whole, and are usually tuned to work with fixed-length testing segments. Furthermore, when more testing data is available, individual phone systems can be combined to improve the discrimination capabilities of the resulting system. Finally, we have shown that it is also possible to complement other acoustic/spectral systems by means of their fusion with the cepstral-trajectories systems.

References

Brummer, Niko and Johan du Preez. 2006. Application-independent evaluation of speaker detection. *Computer Speech and Language (CSL)* 20. 230–275.

Dehak, Najim, Patrick J. Kenny, Reda Dehak, Pierre Dumouchel and Pierre Ouellet. 2011. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing* 19(4). 788–798.

Castro, Alberto de, Daniel Ramos and Joaquin Gonzalez-Rodriguez. 2009. Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking. *Interspeech*. 2343–2346.

Ferrer, Luciana. 2009. *Statistical modeling of heterogeneous features for speech processing tasks*. PhD dissertation: Stanford University.

Gonzalez-Rodriguez, Joaquin, Phil Rose, Daniel Ramos, Doroteo T. Toledano and Javier Ortega-Garcia. 2007. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing* 15(7). 2104–2115.

Kajarekar, Sachin S., Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer and Tobias Bocklet. 2009. The SRI NIST 2008 Speaker Recognition Evaluation System. *International Conference on Acoustics, Speech and Signal Processinng (ICASSP)*. 4205–4209.

Kenny, Patrick. 2010. Bayesian speaker verification with heavy tailed priors. *Keynote presentation at Odyssey 2010*. Brno, Czech Republic.

Kenny, Patrick,Pierre Oullet, Najim Dehak, Vishwa Gupta and Pierre Dumouchel. 2008. A Study of Inter-speaker Variability in Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing* 16(5). 980–988.

Kinnunen, Tomi and Haizhou Li. 2010. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication* 52. 12–40.

National Institute of Standards and Technology. 2006. NIST SRE 2006 Evaluation Plan.http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf (7 March, 2006.)

Reynolds, Douglas A., Thomas F. Quatieri and Robert B. Dunn. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing* 10. 19–41.

Rose, Phil and Elaine Winter. 2010. Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses. *International Conference on Speech Science and Technology (SST).* Melbourne, Australia: Australasian Speech Science and Technology Association.

Shriberg, Elizabeth, Luciana Ferrer, Sachin S. Kajarekar, Anand Venkataraman and Andreas Stolcke. 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication* 46 (3–4). 455–472.

Wikipedia. 2012. Arpabet. *Wikipedia, The Free Encyclopedia*. Available online at: http://en.wikipedia.org/wiki/Arpabet (Accessed 19 July, 2012.)

3.6. Linguistically-constrained formant-based i-vectors for automatic speaker recognition

Título: "Linguistically-constrained formant-based i-vectors for automatic speaker recognition"
Autores: Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez
Revista: Speech Communication
Volume 76 Issue C, February 2016
Páginas: 61-81
Editor: Elsevier Science Publishers B. V. Amsterdam, The Netherlands

Linguistically-constrained formant-based i-vectors for automatic speaker recognition[☆]

Javier Franco-Pedroso*, Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group, EPS, Universidad Autonoma de Madrid, c/Francisco Tomas y Valiente 11, 28049 Madrid, Spain

Abstract

This paper presents a large-scale study of the discriminative abilities of formant frequencies for automatic speaker recognition. Exploiting both the static and dynamic information in formant frequencies, we present linguistically-constrained formant-based i-vector systems providing well calibrated likelihood ratios per comparison of the occurrences of the same isolated linguistic units in two given utterances. As a first result, the reported analysis on the discriminative and calibration properties of the different linguistic units provide useful insights, for instance, to forensic phonetic practitioners. Furthermore, it is shown that the set of units which are more discriminative for every speaker vary from speaker to speaker. Secondly, linguistically-constrained systems are combined at score-level through average and logistic regression speaker-independent fusion rules exploiting the different speaker-distinguishing information spread among the different linguistic units. Testing on the English-only trials of the core condition of the NIST 2006 SRE (24,000 voice comparisons of 5 minutes telephone conversations from 517 speakers -219 male and 298 female-), we report equal error rates of 9.57% and 12.89% for male and female speakers respectively, using only formant frequencies as speaker discriminative information. Additionally, when the formant-based system is fused with a cepstral i-vector system, we obtain relative improvements of ∼6% in EER (from 6.54% to 6.13%) and ∼15% in minDCF (from 0.0327 to 0.0279), compared to the cepstral system alone.

Keywords: automatic speaker recognition; formant frequencies; formant dynamics; linguistically-constrained systems

1. Introduction

Most of the studies in automatic speaker recognition over the last two decades have been based on compact representations of the speech signal in short analysis windows (i.e.

[☆]Non-standard abbreviations: NIST: US National Institute of Standards and Technology. SRE: Speaker Recognition Evaluation. ASR: Automatic Speech Recognition.

*Corresponding author.

Email addresses: javier.franco@uam.es (Javier Franco-Pedroso), joaquin.gonzalez@uam.es (Joaquin Gonzalez-Rodriguez)

MFCC, RASTA-PLP, etc.) [1]. Although they are based on spectral representations of the speech signal, it is difficult to directly relate the physiological traits of an individual with the set of such extracted features due to the additional transformations to which they are subjected (inverse FFT, DCT, etc.) [2]. Moreover, it is hard to interpret such kind of coefficients inasmuch as they do not correspond to any physical magnitude but to mathematical abstractions (the so-called cesptral domain). Formant frequencies, on the other hand, represent the resonant frequencies of the vocal tract of an individual, being easily interpretable and directly related with anatomical and physiological characteristics [3] [4]. This makes them specially suitable for forensic purposes [5] [6], where formant measurements have been used for forensic voice comparison for several decades [7] [4].

Voice comparison is usually performed in the context of linguistic units in forensic-phonetics [8], but reported studies are usually based on limited experimental frameworks (in terms of number of speakers, number of analysed linguistic-units, or both) due to the manual processes involved in order to extract formant frequencies or labelling the analysed units. So, it is of broad interest to analyse the abilities of formant frequencies for speaker recognition following a similar approach but applied on a large-scale experimental framework with the aid of fully automatic systems. In this way, the presented results can give useful insights for the practitioners in that field.

Furthermore, interpretable features are helpful in order to correlate with human observations and may lead to find some clues that could be hidden even for very complex cepstral-based systems [9]. Such kind of interpretable features, or the systems that make use of them, are usually classified as *higher-level* [10], and sometimes involve some kind of *constraints* [11] that are applied either in the feature extraction process (in order to define the feature itself), in the speaker modelling process (in order to reduce the intra-speaker variability), or both of them [10]. *Higher-level* systems provide very useful and complementary information that usually leads to performance improvements when they are combined with short-term acoustic systems [12] [13] [14].

With the objective of using interpretable features as formant frequencies but being able to evaluate them in the same challenging conditions of the state-of-the-art systems (e.g. the NIST Speaker Recognition Evaluations framework), we present in this paper a speaker verification system based on formant frequencies through the combination of different linguistically-constrained i-vector systems. While previous approaches [12] [15] [16] [17] extract the speaker distinguishing information from formant frequency dynamics through trajectories coding in the context of some linguistic units (phones, diphones, syllables or pseudo-syllables), in this work we address this issue by means of the classical derivative coefficients [18] [19], also known as *delta* (Δ) features, widely used in speech processing [20] in order to account for the dynamic information in the cepstral domain. This approach has the advantage of not reducing each linguistic segment (e.g. phone, diphone, *etc.*) to a single observation vector, relaxing the previous requirements of training data derived from extracting one single feature vector per linguistic segment.

The rest of the paper is organized as follows. Section 2 presents a brief overview of how formant frequencies have been used for speaker recognition, while Section 3 describes the automatic feature extraction process followed in the proposed approach. Section 4 details

how linguistically-constrained i-vector systems are built from formant features with the aid of automatically-generated phonetic labels. Section 5 describes the constraint-selection rules and fusion techniques used in order to combine the linguistically-constrained systems for text-independent speaker recognition. The experimental framework and evaluation metrics are presented in Section 6, including a description of our reference cepstral-based speaker recognition system. Results are shown in Section 7 for both independent linguistically-constrained systems and for several constraint combinations, as well as for the combination of formant and cepstral-based systems. Finally, conclusions are drawn in Section 8 and extended results are reported in a final appendix.

2. Formant frequencies for speaker recognition

Formant frequencies have strong individualization potential [7] and have been used for forensic voice comparison for several decades [4]. Usually, formant centre frequencies are extracted at the temporal midpoint of vowels [21] reflecting in part certain anatomical dimensions of a speaker as the length and configuration of the vocal tract. Also, the mean frequencies over the time-course of the vowel [22] have been used.

In order to obtain richer representations, frame-by-frame formant-frequency distributions have been modelled through either long-term formant distributions (LTFs) [3] or multivariate Gaussian mixture models (GMMs) [5]. It is also common to incorporate formant bandwidth measurements in order to complement the information provided by instantaneous formant frequency values [5] [16], as they are also related to vocal tract conditions.

Formant dynamics were also proposed for speaker recognition [8] under the assumption of presenting higher inter-speaker variability within linguistic units than the static measurements of formant frequencies: while speakers seems to show very similar acoustic properties at moments at which 'phonetic targets' [8] are achieved (e. g. formant frequencies at a segment's temporal midpoint), much larger differences are exhibited in the ways they move between consecutive targets [23].

This transitional information is omitted by statistical distributions obtained from frame-by-frame formant frequencies. In order to capture this dynamic information, two main approaches have been used: polynomial fitting [8] [12] and Discrete Cosine Transform (DCT) [24] [17] of formant trajectories over linguistic units. Both approaches compute a fixed number of polynomial or DCT coefficients per trajectory and concatenates the coefficients from the different formant trajectories, yielding a single feature vector that captures the dynamic information of the different formants in a given linguistic unit. In order to define the speech region where formant trajectories are computed, both manual segmentations (mainly in the forensic field) [24] [8] and automatic speech recognition (ASR) systems [12] [17] have been used. Using coded trajectories as feature vectors, speakers have been modelled through multivariate kernel distributions (MVK) [24] [16] or GMM's [17] in a linguistic unit-dependent manner, or by means of joint factor analysis (JFA), compensating for intersession variability, by pooling together trajectories from different units [12].

Similarly, the approach proposed in this paper is based on formant frequencies, but extracts the dynamic information through derivative coefficients [18] [19] regardless of the lin-

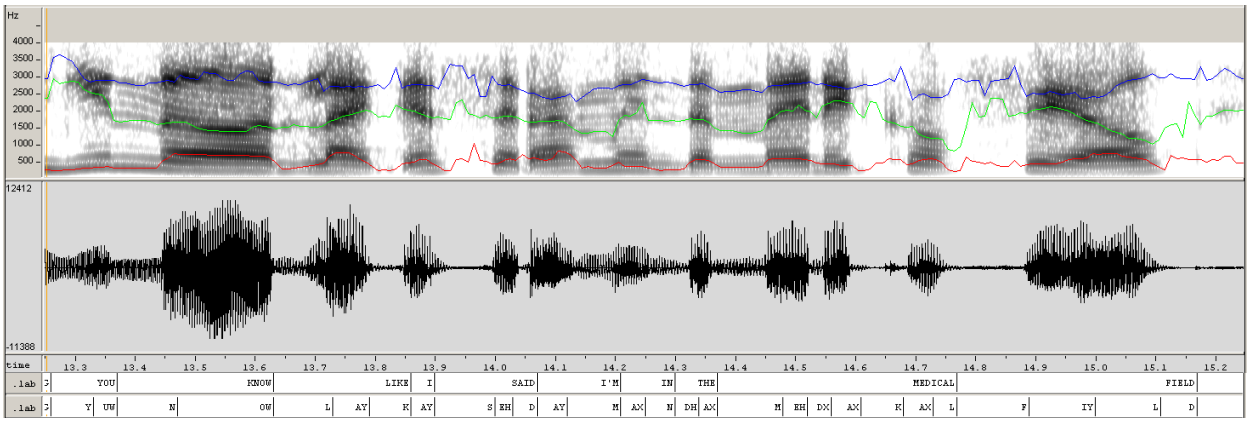


Figure 1: *Formant tracking for a speech sample in Wavesurfer©, and its corresponding automatically extracted word and phone labels.*

guistic content. These coefficients are also extracted at a frame-by-frame rate and combined with the static information of instantaneous formant frequency values. Then, linguistic units are used as constraints applied to feature vectors in order to develop separate i-vector systems for each linguistic unit, allowing to independently analyse their speaker-distinguishing abilities.

3. Feature extraction

3.1. Formant tracking

Several methods and algorithms have been proposed for formant tracking [20], but only some of them have been implemented and made available within free software packages, as for example Wavesurfer [25], Praat [26] or WinSnoori [27]. Among them, the first one was selected for this work because it allows to easily automate this process for large databases. Wavesurfer is a general-purpose software audio editor widely used for studies of acoustic phonetics that provides an interactive display for waveform, spectrograms, pitch tracks or transcriptions visualization, therefore being a graphical user-oriented tool. However, it's developed using the Snack Sound Toolkit library [28], so scripts for automatic processing of large databases can be written in Tcl/Tk [29].

The Snack formant tracker bases its formant-frequency estimates on a linear prediction analysis performed at each frame, and dynamic programming is used to refine the resulting trajectories [30]. It was used with default parameters for both male and female speakers, except for the number of formant frequencies to be tracked. Most formant tracking estimators focus on formants F_1-F_3 due to the fact that higher formants are progressively weaker in intensity [20]. Moreover, the average frequency position of F_4 is 3500 Hz, which is close to the cut-off frequency of the telephone-line band-pass filter. As in this work we are dealing with telephone-line speech, formant frequencies have then been extracted for the first three formants, with a 10 ms time resolution.

For the sake of simplicity in the feature-extraction phase, and due to the large number of speakers and linguistic units present in our experimental framework, no specific settings were

used for different speakers or units but a common one. For similar reasons, no exhaustive analysis was made regarding the suitability of the settings used, but just some shallower checks against typical formant values for the measurements obtained. As an automatic system, it will present errors that the following stages have to deal with.

3.2. Dynamic-information

While frame-by-frame formant frequencies can be estimated regardless of the linguistic content present in the speech signal, formant trajectories, as they have been used so far in speaker verification [12] [16] [17], can only be defined by using phonetic segmentations in order to delimit the speech region on which they are going to be coded. Working with automatic systems, both formant tracking errors and misalignment of phone label from the ASR will be observed, leading to erroneous coded trajectories in those cases.

An example is shown in Figure 2, where the phonetic transcription is correct but not properly aligned with the beginning of the acoustic signal and, therefore, spurious formant values computed at the beginning of the segment give rise to an artificial trajectory. If, for example, polynomial fitting is used in order to code the trajectory, the artificial spiky trajectory will require larger values in the higher order coefficients. Thus, the single feature vector corresponding to the whole linguistic unit will provide misleading information. Also, the same problem appears if some isolated spurious formant values arise within a well aligned phonetic transcription. On the contrary, if a frame-by-frame feature-extraction scheme is followed as will be used here, isolated spurious formant values only affect to the feature vectors extracted in these frames instead of the whole linguistic segment.

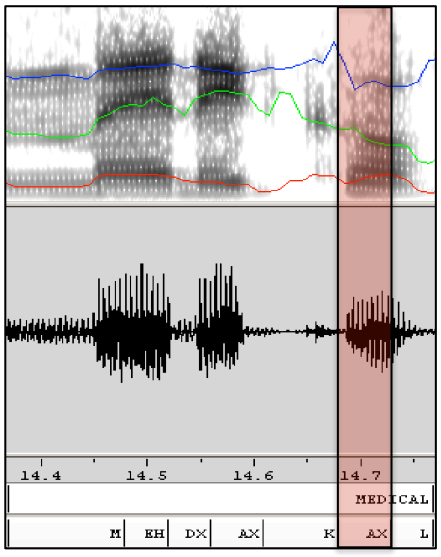


Figure 2: *Example of label temporal misalignment.*

Moreover, although the one-vector-per-linguistic-segment coding approach achieves a highly compact representation of formant trajectories in linguistic units, it greatly reduces the amount of data that can be used to train the parameters of the system, specially for

linguistic units with low frequency of occurrence. This problem is aggravated when linguistic constraints are applied for speaker modelling and comparison. For example, the diphone units analysed in this work present, on average, a frequency of occurrence of 10 times per conversation. Thus, if the trajectories of the first three formant frequencies are coded with the first 5 coefficients of its DCT, and the coefficients concatenated in a single feature vector, only ten 15-dimensional (3 formant trajectories $\times$ 5 coefficients/trajectory) feature vectors will be available per conversation. Thus, sufficient statistics for the speaker modelling process have to be computed from a reduced number of observations (even lower than the number of features per observation). However, if a frame-by-frame feature extraction scheme is followed instead, the larger length of diphones will provide enough number of feature vectors in order to extract reliable sufficient statistics even with a low number of occurrences per conversation. In the previous example, assuming an average number of samples per diphone equal to 10, one hundred 3-dimensional samples will be available.

For these reasons, the *delta* (Δ) or derivative coefficients have been used to account for the dynamic information of formant frequencies instead of trajectory coding. Although delta coefficients cannot include the whole formant trajectory along the linguistic segment, they can characterize the local dynamic information while keeping a frame-by-frame feature extraction scheme. Delta coefficients were originally introduced for cepstrum coefficients [18] [19] in order to characterize the spectral transitional information, and are part of typical state-of-the-art speaker recognition systems. Applied to formant frequencies, this time derivative, approximated by a finite difference, has the following form

$$\frac{\delta F_m(t)}{\delta t} \approx \Delta F_m(t) = \frac{\sum_{k=-K}^{K} k h_k F_m(t+k)}{\sum_{k=-K}^{K} h_k k^2} \tag{1}$$

where $F_m(t)$ is the m-th formant frequency at time t and h_k is a window of length $2K+1$ frames. In this study, a rectangular window ($h_k = 1$) is used with $K = 2$.

Finally, derivative coefficients are appended to instantaneous formant frequencies for each frame, giving rise to our 6-dimensional feature vectors at frame resolution (10 ms), $f(t)$.

$$f(t) = [F_1(t), F_2(t), F_3(t), \Delta F_1(t), \Delta F_2(t), \Delta F_3(t)] \tag{2}$$

While additional dynamic information could be added in a similar setting through the *delta-delta* (or *acceleration*) coefficients [20], this option has been discarded for practical reasons. As it will be shown in the following Section, independent speaker recognition systems are developed based on the different linguistic constraints. Thus, the number of feature vectors available for developing each independent system is highly reduced due to the region-conditioning process. If the dimensionality of the feature vectors is further increased, the ratio between the number of training samples and the complexity of the models is further reduced. As a trade-off between the amount of information and the complexity of the models, only delta features have been included in order to account for the dynamic information of formant frequencies.

Similarly, formants bandwidth information, while also used in forensic voice comparison, has been discarded based on preliminary experiments where including both formant frequencies and bandwidths did not improve the average performance across the different constraints, and have not been considered for further experiments in this work.

4. Linguistically-constrained speaker verification

Linguistically-constrained systems make use of an automatic speech recognition (ASR) system in order to condition the speech regions to be processed. ASR conditioning has been applied in automatic speaker recognition systems based on both cepstral [11] and higher-level [32] features. For cepstral-based systems, ASR conditioning is applied after the feature extraction process, defining the *constraints* to be applied by each subsystem to the features that can be used in speaker modelling and comparison stages. In this way, the intra-speaker variability due to the different lexical content between training and testing utterances is reduced. In the case of higher-level features, constraints are needed in order to define the feature itself, as they usually attempt to capture the dynamic behaviour of a specific measurement (pitch, energy, *etc.*) over several speech frames [10]. This is also the case of formant trajectories coding in the context of linguistic units. However, for systems based on prosodic information, once the features have been extracted, features belonging to different linguistic units are usually pooled together [12] [13] for the speaker modelling and comparison stages.

In this work, although ASR conditioning is avoided in the feature extraction process, constraints are applied in the speaker modelling and comparison stages. In this way, we aim not only to reduce the intra-speaker variability but also to test the discriminative abilities of formant frequencies within each linguistic unit independently, which can provide useful insights, specially to practitioners in forensic phonetics. Moreover, this allows to adopt a flexible approach to automatic speaker recognition where the linguistic specificities of particular speakers can be taken into account by using speaker-dependent constraints.

With this objective, we have developed independent i-vector systems [33] for each of the linguistic constraints under analysis, running in parallel for each speaker comparison (or *trial* in NIST SREs nomenclature) over the set of features belonging to its corresponding constraint. Additionally, calibrated likelihood-ratios (LRs) from a given subset of constraints can be combined in order to provide a single LR per trial.

4.1. Region conditioning

For the purpose of automatic region conditioning, we use the labels provided by an automatic speech recognition (ASR) system that produces transcriptions defining both phonetic content and time interval of speech regions in which the audio stream can be segmented. In this work, the phonetic transcription labels produced by the SRIs Decipher state-of-the-art ASR system [34] are used. For this system, trained on English data from telephonic conversations, the Word Error Rate (WER) on native and non-native speakers on the transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and

Vowels					
Monophthongs			*Monophthongs*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
AO	ɔ	off; fall; frost	AE	æ	at; fast
AA	ɑ	father; cot	*Diphthongs*		
IY	i	bee; see	Arpabet	IPA	Word examples
UW	u	you; new; food	EY	eɪ	say; eight
EH	ɛ	red; men	AY	aɪ	my; why; ride
IH	ɪ	big; win	OW	oʊ	show; coat
UH	ʊ	should; could	AW	aʊ	how; now
AH	ʌ	but; sun	*R-coloured vowels*		
	ə	sofa; alone	Arpabet	IPA	Word examples
AX		discus	ER	ɜ	her; bird; heart; nurse

Consonants					
Stops			*Affricates*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
P	p	pay	CH	tʃ	chair
B	b	buy	JH	dʒ	just
T	t	take	*Semivowels*		
D	d	day	Arpabet	IPA	Word examples
K	k	key	Y	j	yes
G	g	go	W	w	way
Fricatives			*Liquids*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
F	f	for	L	ɬ	late
V	v	very	R	ɹ *or* ɻ	run
TH	θ	thanks; Thursday	DX	ɾ	wetter
DH	ð	that; the; them	*Nasals*		
S	s	say	Arpabet	IPA	Word examples
Z	z	zoo	M	m	man
SH	ʃ	show	N	n	no
HH	h	house	NG	ŋ	sing

Table 1: *39 phones from the Arpabet phonetic transcription code and their correspondent IPA symbols (extracted from [31]).*

36.1% respectively. While these results are equivalent to those obtained by other state-of-the-art systems on similar databases [35], transcription errors will be non negligible and will produce that, in order to compute the i-vector for a particular linguistic unit, some frames belonging to a different one will be taken into account, degrading the performance of the system based on that unit. In this work, no exhaustive analysis has been done regarding whether the errors occurred are associated with particular units or speakers, as we have no transcriptions available for the datasets used.

An analysis of such kind can be found in [36], where it is shown that errors are related with "*extreme prosodic characteristics, words occurring turn-initially, as discourse makers or preceding disfluent interruption points, and acoustically similar words that also have similar language model probabilities*". Thus, errors seem not to be associated with specific units but influenced by several aspects. It is also highlighted that "*speaker differences cause enormous variance in error rates*", and this seems to be "*not fully explained by differences in word choice, fluency, or prosodic characteristics*". Thus, a plausible cause can be the acoustic specificities of different speakers.

Regarding the results reported in this work, on one hand, a variable ASR performance across units would affect the relative performance among systems based on them. Thus, if a particular unit present worse speaker recognition performance than other, this can be due not only to a less discriminative ability of its formant frequencies but also to the fact that more ASR errors may occur for that particular unit. On the other hand, a variable ASR performance across speakers will reflect, in fact, the particularities of the different speakers, which will be combined with the different discriminative abilities of formant frequencies.

4.2. Types of constraints

Looking for multiple separate contributions to the speaker identity in a speech file, linguistic units are the natural and straightforward group of segments to work with. ASR labels allow to define a large set of candidate constraints from linguistic units [16], showing each of them different characteristics in terms of within-unit formant dynamics, unit-length and frequency of occurrence. Among them, the following were used:

- Phones: although they are the shortest units and can appear in many different linguistic contexts, their high frequency of occurrence allow to develop more robust constrained systems. For this work, 39 phone units from an English lexicon plus two filled pauses (represented as PUH and PUM) were selected. These linguistic units are represented by the "2-character" ARPABET symbols [37] in the phonetic transcriptions provided by the ASR system [34]. Table 1 shows the correspondence between Arpabet symbols and the International Phonetic Alphabet (IPA) ones, while Figure 3 shows an example of region conditioning for a particular phone unit.

- Diphones: defined as every possible combination of phone pairs, the 98 most frequent diphones were selected. Compared with phones, they present longer length but much lower frequency of occurrence. However, they show less contextual variation, which may lead to reduce the intra-speaker variability of formant dynamics between different occurrences of the same diphone.

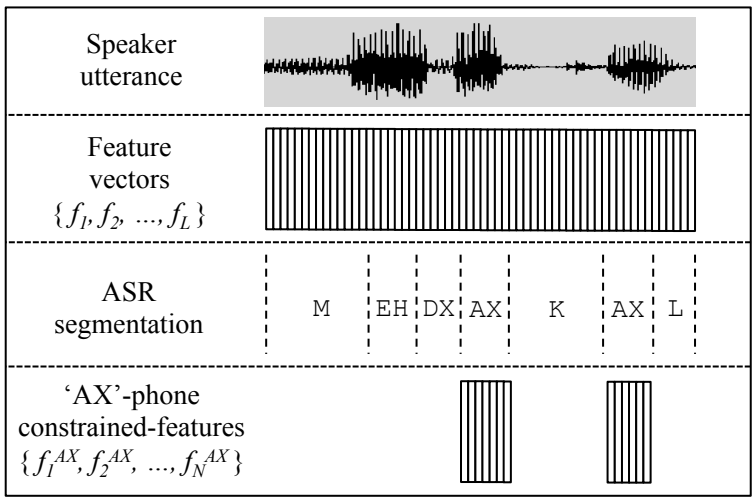


Figure 3: *Example of region conditioning for a particular phone unit ('AX').*

Although formants obtained from consonants are not regularly analysed within phonetics, we have not restricted the analysis to some specific units (e.g. only vowels, or vowels and some voiced consonants) for two main reasons. First, as the authors are neither linguists nor phoneticians but engineers, the only restriction applied regarding the linguistic units to be analysed is that they present enough frequency of occurrence. And secondly, working with a wide range of linguistic units illustrates the power of using automatic systems, providing a thorough analysis of their individualization potential.

4.3. Linguistically-constrained i-vector systems

An i-vector system [33] is a factor analysis (FA) based front-end for speaker verification which attempts to summarize the speaker distinguishing information in a given utterance, represented by a set of L feature vectors $\{f_1, f_2, ..., f_L\}$, through a single low-dimensional vector, the so-called identity vector or *i-vector* for short. This i-vector w accounts for the speaker and channel/session information present in a given utterance, representing it in a low-dimensional variability subspace. This is done converting the speaker- and session-independent supervector (m), usually taken to be the UBM supervector, into the speaker- and session-dependent supervector (M) that represents a given speaker utterance:

$$M = m + Tw \tag{3}$$

where T is a rectangular matrix of low rank defining the total variability (TV) space that contains the speaker and channel variability. For the purpose of developing linguistically-constrained systems, this FA model is applied in this work for every given constraint, C:

$$M^C = m^C + T^C w^C \tag{4}$$

Thus, independent UBMs and TV subspaces are trained on the background dataset (see Section 6 for details) from every linguistically-constrained set of feature vectors $\{f_1^C, f_2^C, ...\}$,

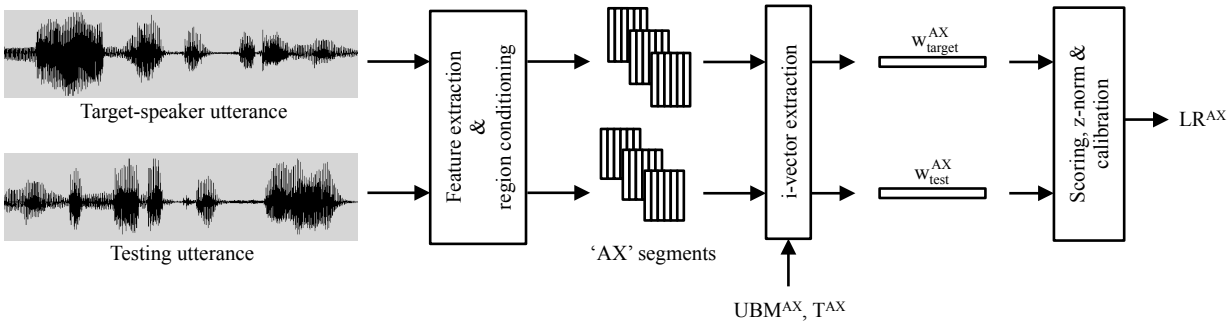


Figure 4: *Linguistically-constrained speaker verification system for a particular phone unit ('AX').*

allowing to obtain a constraint-dependent i-vector (w^C) from the occurrences of a given unit within an utterance (see Figure 4). Both the number of components of the UBM (ranging from 2 to 256) and the number of dimensions of the TV space (ranging from 5 to 50) are optimized on the development dataset (see Section 6 for details) for each linguistic-unit/constraint. Extracted i-vectors are length-normalized and whitened [38] previously to the scoring stage. Then, for a given constraint C, the similarity measure (score) between the target speaker (w^C_{target}) and the testing utterance (w^C_{test}) i-vectors is given by the cosine distance between them:

$$score(w^C_{target}, w^C_{test}) = \frac{\langle w^C_{target}, w^C_{test} \rangle}{\|w^C_{target}\| \|w^C_{target}\|} \tag{5}$$

Finally, constraint-dependent scores are z-normalized [39] and calibrated in an application-independent way [40] through *logistic regression* trained on the development dataset, using the FoCal toolkit [41].

5. Combination of linguistically-constrained systems

For a given speaker comparison, the final likelihood ratio can be either any of the constraint-dependent ones or a combination of a subset of them. In this Section, several strategies have been followed, regarding different aspects, in order to tackle the issue of how to combine the different linguistic constraints. First, the type of linguistic constraints taken into account has to be set. Then, some rule must be followed in order to select the particular constraints to be fused, according to some criterion. Finally, a specific fusion technique must be used in order to combine the likelihood ratios corresponding to different constraints.

5.1. Linguistic-constraint types

In a first stage, constraint combinations have been analysed separately for phone and diphone units. As diphone units are defined as two-phone combinations, they share the same information as phones, but spread over different diphones. However, dynamic information of the transition between two specific phone units is only modelled by diphone units, which may provide significant discrimination ability between speakers. Finally, constraint combinations

will be analysed when pooling together both types of linguistic units in order to test if the transitional information provided by diphone units provide additional discrimination ability to phone units.

5.2. Constraint-selection rules

We address the issue of constraints selection to be fused as a feature selection process [42], testing two constraint-selection schemes as in [17]:

- N-best performing units: for this method, constraints are sequentially fused in decreasing performance order on the development dataset. Once the EER is known for every number of constraints to be fused (see Figure 12a), the subset of constraints with the best performance on the development dataset is selected and applied in the evaluation dataset.

- Sequential Forward Selection (SFS): similarly to the previous method, constraints are sequentially fused in decreasing performance order on the development dataset. However, instead of keeping every subsequent constraint, they are included into the fusion subset only if the performance of the fused system increases. This procedure can be summarized in the following steps:

  1. Take the best-performing constraint as the initial subset.
  2. Take the next best-performing constraint and fuse with the previous subset. If the performance of the fused system is increased with respect to that of the previous step, keep the constraint; otherwise, reject it.
  3. Repeat the previous step until the worst performing constraint is reached.

5.3. Fusion techniques

Two different fusion techniques have been analysed in this work. First, a simple fusion rule consisting on averaging the log-LRs of the subset of N constraints to be combined has been applied through

$$logLR = \frac{1}{N} \sum_{\forall C \, in \, subset} logLR^C \tag{6}$$

where ($logLR^C$) is the log-LR for a particular constraint C. While this technique do not take into account the different performance of the different constraints, it has the advantage of not requiring additional training data.

Secondly, a linear combination of log-LRs is applied through

$$logLR = \alpha_0 + \sum_{\forall C \, in \, subset} \alpha^C logLR^C \tag{7}$$

where the vector of weights $\boldsymbol{\alpha} = [\alpha_0, \alpha^{C_1}, \alpha^{C_2}, ..., \alpha^{C_N}]$ is obtained by *logistic regression* [43] training on the development database, using the FoCal toolkit [41].

For both fusion techniques, missing trials are handled in the same way as in [44]. Missing trials may appear when the corresponding constraint is not present in either target-speaker or testing utterances. In such cases, the corresponding sub-system cannot contribute a log-LR for that trial. However, as every linguistically-constrained system is independently calibrated, log-LRs of zero are inserted for missing trials in order to have valid log-LRs for every sub-system to train the fusion rule.

6. Experimental framework

One of the main goals of this work is to quantify the discriminative power of formant frequencies and their dynamics on the experimental frameworks used by the automatic speaker recognition community. NIST SREs have become a *de facto* standard for testing automatic speaker recognition systems, providing since 1997 [45] increasingly challenging datasets and protocols.

In order to develop and test the proposed speaker verification systems, we have used the datasets and protocols belonging to the NIST SREs carried out on years 2004 [46], 2005 [47] and 2006 [48], mainly those corresponding to the *core conditions*, which are composed of 5-minutes length telephone-line recordings of conversational speech. Among them, only English conversations have been used in order to match the characteristics of the ASR system [34].

Two are the main reasons for using only those years NIST SREs. First, the authors have access only to the ASR labels corresponding to those datasets, kindly provided by SRI. And second, the core condition of the NIST 2006 SRE is the main evaluation benchmark where a high number of comparative results are available from different high-level systems [10] [49] [12] [15] [13].

6.1. Performance evaluation metrics

The main evaluation metric used along this work to measure the discriminative performance is the equal error rate (EER) [45]. It is also used as the criterion by which the subsets of constraints are selected for the combination of systems. However, in accordance to the protocols used [48], the minimum of the C_{Det} (minDCF) is also shown. Finally, the C_{llr} cost function and the calibration loss (C_{llr}^{loss}) [40] are included as well in order to evaluate the calibration properties [50] of the different constraints and fusion schemes.

6.2. Background, development and evaluation datasets

The experimental protocol has been carefully designed in order to avoid obtaining overoptimistic results due to any overlap between datasets belonging to different development stages. With this aim, different datasets have been devoted to different purposes.

- Background: NIST 2004 SRE dataset [46] has been used as the background dataset for training UBMs and total variability matrices. This dataset comprises 2,541 files (1378 5-minutes, 581 30-seconds and 582 10-seconds long) from 125 male speakers and 3,626 files (2022 5-minutes, 802 30-seconds and 802 10-seconds long) from 187 female

speakers. Also, speakers cohorts for Z-normalization were extracted from this dataset, using one 5 minute recording per speaker.

- Development: NIST 2005 SRE dataset [47] has been devoted to perform parameter optimization of the systems. Target speakers from the 1side-1side task were divided into two halves in order to have two different testing frameworks: *sre05-cal* and *sre05-val*, consisting both of them in ∼5,500 male trials from ∼120 target speakers and ∼7,400 female trials from 171 target speakers. The number of both UBM components and dimensions of the TV subspace were optimized by minimizing the EER bias and variance over these two testing frameworks. Once the parameters of the system for each constraint were set, scores from *sre05-cal* were used to train the calibration process (logistic regression) and scores from *sre05-val* to train the fusion schemes.

- Evaluation: English-only trials from the core condition of the NIST 2006 SRE [48] were used for evaluating the proposed approach, consisting of 9,720 male trials for 219 target speakers and 14,293 female trials for 298 target speakers.

6.3. Reference system

Our cepstral-based reference system is also an i-vector system developed by using the same experimental framework as the linguistically-constrained formant-based systems. It is based on mean-normalized, RASTA-filtered and gaussianized MFCC features (19 coefficients plus deltas). 1024-component UBMs and 600-dimensional TV subspaces were trained for each gender. Unlike for the formant-based system, LDA (trained on the background dataset) was applied in order to compensate for the intersession variability [33]. Thus, the similarity measure (score) between a target speaker (w_{target}) and a testing utterance (w_{test}) i-vectors is given by

$$score(w_{target}, w_{test}) = \frac{(A^t w_{target})(A^t w_{test})}{\sqrt{(A^t w_{target})(A^t w_{target})}\sqrt{(A^t w_{test})(A^t w_{test})}} \tag{8}$$

being A the LDA matrix. Finally, scores are z-normalized and calibrated in the same way as the linguistically-constrained systems.

7. Results

7.1. Independent linguistically-constrained systems

7.1.1. Overall performance per constraint

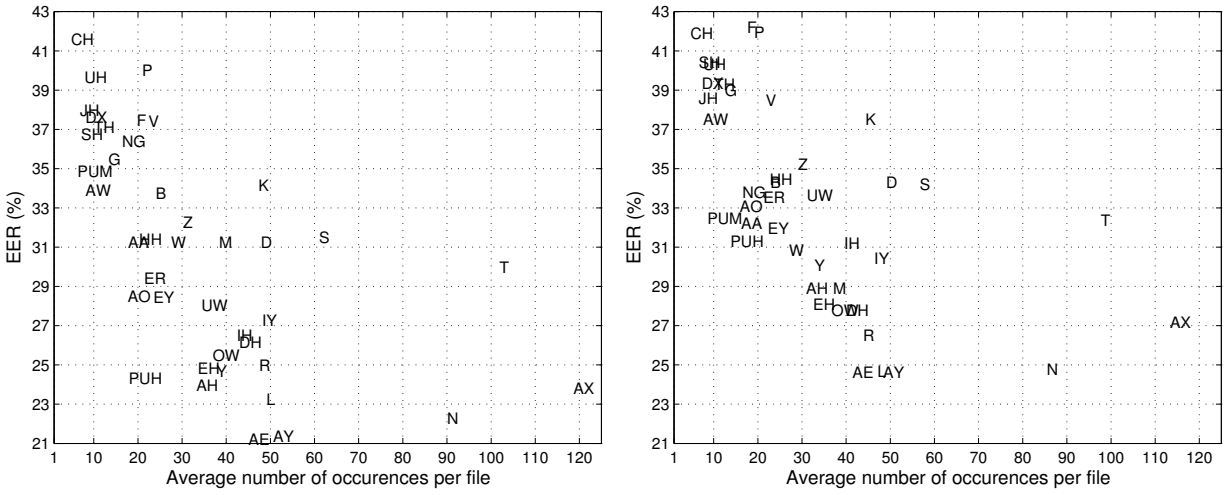
In this section we show the performance of each linguistically-constrained system independently. Table 2 shows the result for each metric on the evaluation dataset for the 10 best-performing phone-constraints (results for every phone are given in Table A.1), while Figure 5 shows the EER as a function of the frequency of occurrence for each of the 41 analysed phone-constraints. In both cases male and female trials are independently analysed; it can be seen that the constraints show similar behaviour for both genders in relative terms

NIST 2006 SRE, English-only trials									
Male					Female				
Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AE	21.21	0.0850	0.6668	0.0143	AY	24.59	0.0841	0.7101	0.0111
AY	21.38	0.0825	0.6580	0.0158	AE	24.59	0.0876	0.7308	0.0131
N	22.26	0.0812	0.6896	0.0168	L	24.68	0.0869	0.7355	0.0127
L	23.24	0.0839	0.7083	0.0133	N	24.77	0.0839	0.7256	0.0112
AX	23.80	0.0844	0.7001	0.0150	R	26.49	0.0932	0.7681	0.0132
AH	23.96	0.0964	0.7286	0.0158	AX	27.15	0.0932	0.7764	0.0100
PUH	24.32	0.0933	0.7296	0.0137	OW	27.79	0.0936	0.7830	0.0098
Y	24.68	0.0915	0.7325	0.0180	DH	27.79	0.0940	0.7876	0.0114
EH	24.83	0.0972	0.7544	0.0140	EH	28.06	0.0990	0.8196	0.0157
R	24.96	0.0937	0.7380	0.0149	AH	28.89	0.0974	0.8185	0.0079

Table 2: *Results on the evaluation dataset for the 10 best-performing phone-constraints (extended results for every phone are given in Table A.1).*

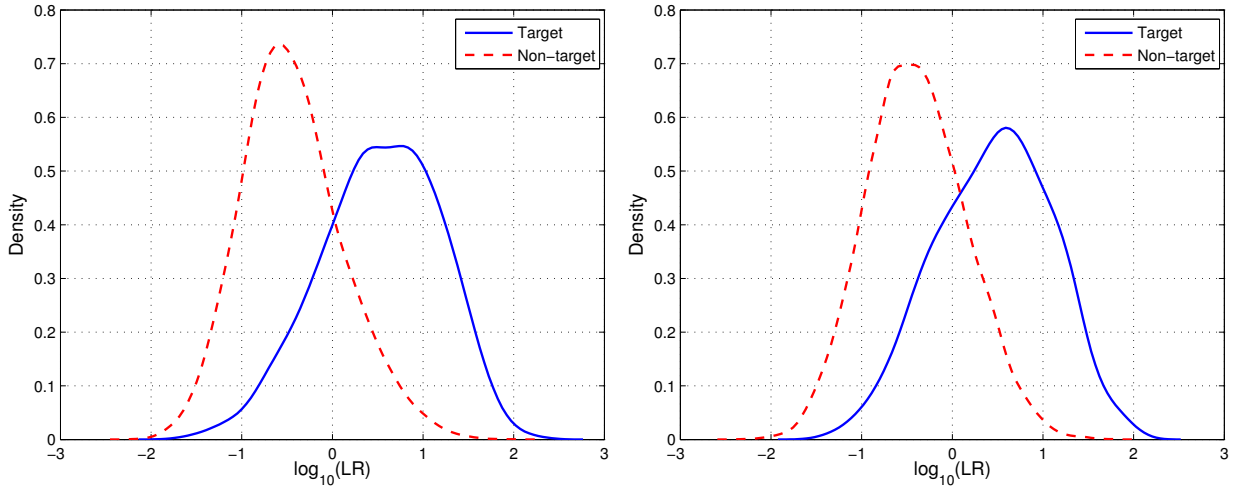
(Figure 5) except for the shift in absolute performance in favour of male speakers, which has been also reported in NIST SRE frameworks for cepstral-based systems [33].

It can be seen from Table 2 that, while each of the constraints have limited discriminative performance by themselves, they have good calibration properties. As an example, probability density functions of the logLRs provided for the best-performing phone-constraint



(a) Male trials (b) Female trials

Figure 5: *EER vs frequency of occurrence for phone-constraints on the English-only trials of the core condition of the NIST 2006 SRE. Detailed frequency of occurrence in Table B.1.*



(a) AE phone-constraint, male trials (b) AY phone-constraint, female trials

Figure 6: *Target and non-target log_{10}-LRs probability density functions for the best phone-constrained system on the English-only trials of the core condition of the NIST 2006 SRE.*

are shown in Figure 6. This allows to obtain informative calibrated LRs for voice comparisons from isolated linguistic units, as it has been suggested by some forensic-phonetics practitioners [24] [4].

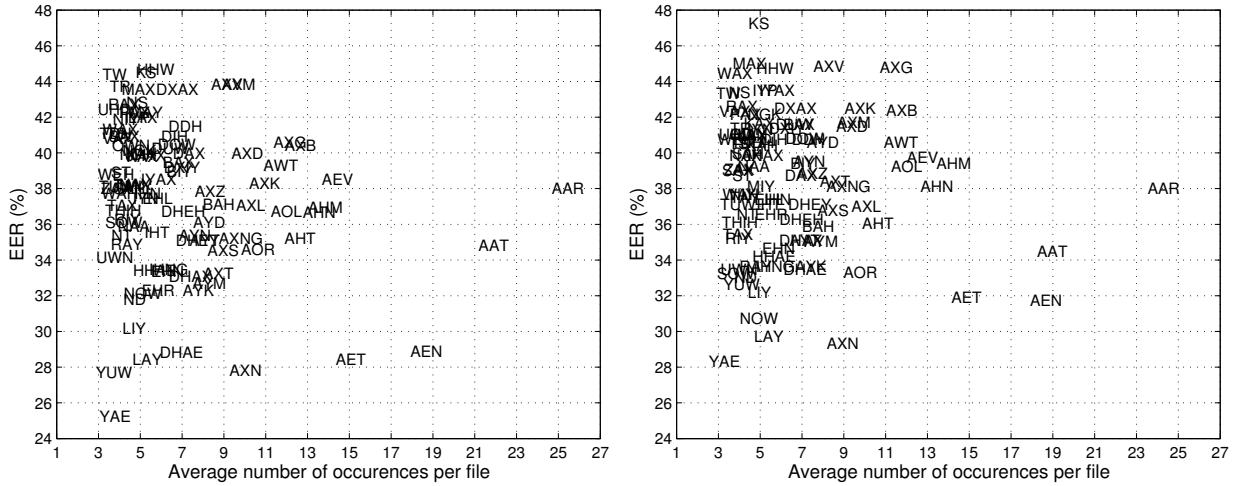
Regarding the relationship between discriminative abilities and frequency of occurrence of each phone-constraint (Figure 5), there is a clear relationship between them in general terms, obtaining lower EERs those constraints with higher frequency of occurrence. However, for a subset of phone-constraints with similar frequency of occurrence, the range of EERs obtained may be wide, suggesting that different linguistic units present different discriminative abilities. In fact, some of the best performing units ('AE', 'AY', 'L', 'R') are not among those with the highest frequency of occurrence. However, it should be noted at this point that neither the formant tracking nor the ASR are error-free processes, and some particular phone units may present more errors than others, affecting to the relative difference in performance among them.

In the case of diphone-constraints, it can be seen from Table 3 that the best performing diphone-constraints are those combining some of the best performing phones (results for every diphone are given in Table A.2). This is a consequence of the combination of instantaneous frequency values with the derivative coefficients, which do not characterize the formant dynamics along the whole unit but in a local vicinity. However, there is not a clear relationship between performance and frequency of occurrence (Figure 7) unlike for phone-constraints, being in fact the best performing constraint, Y-AE, one of the least frequent in the database. This suggests that there is significant speaker-distinguishing information in formant dynamics in the transition between Y and AE phones: although these isolated phone-constraints are two of those with better performance, other two phone combinations among the 10-best performing phone-constraints obtain lower performance despite having a higher frequency of occurrence (e.g. AE-N or AX-N). However, it can be seen that, in aver-

NIST 2006 SRE, English-only trials									
Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
Y-AE	25.26	0.0867	0.7533	0.0245	Y-AE	28.32	0.0894	0.7960	0.0130
Y-UW	27.69	0.0928	0.8005	0.0135	AX-N	29.34	0.0942	0.8190	0.0154
AX-N	27.79	0.0947	0.7829	0.0132	L-AY	29.71	0.0943	0.8154	0.0199
AE-T	28.42	0.0987	0.8103	0.0195	N-OW	30.71	0.0957	0.8489	0.0159
L-AY	28.42	0.0957	0.8215	0.0277	AE-N	31.71	0.0962	0.8528	0.0097
DH-AE	28.82	0.0949	0.8269	0.0171	AE-T	31.90	0.0997	0.8614	0.0128
AE-N	28.88	0.0945	0.8132	0.0129	L-IY	32.20	0.1000	0.8728	0.0114
L-IY	30.15	0.0966	0.8331	0.0130	Y-UW	32.65	0.0948	0.8667	0.0114
N-D	31.80	0.0974	0.8469	0.0152	N-D	33.00	0.0978	0.8736	0.0082
N-OW	32.15	0.0959	0.8665	0.0128	S-OW	33.22	0.0996	0.8811	0.0116

Table 3: *Results on the evaluation dataset for the 10 best-performing diphone-constraints (extended results for every diphone are given in Table A.2). Sample words for listed diphones are: yeah (Y-AE), you (Y-UW), second (AX-N), at (AE-T), like (L-AY), that (DH-AE), an (AE-N), firstly (L-IY), and (N-D), know (N-OW), so (S-OW).*

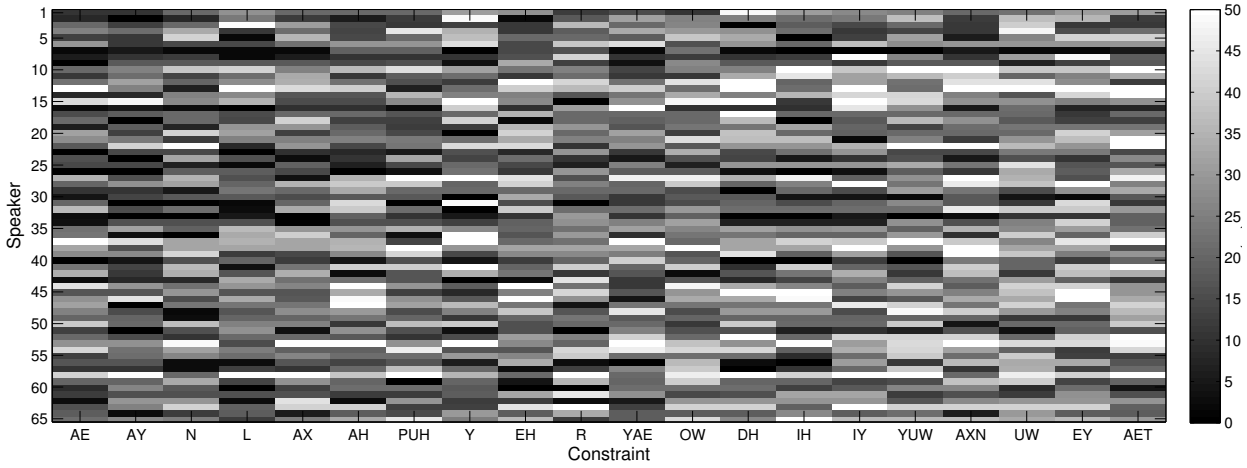
age, diphone-constraints are less discriminative than phone-constraints due to their smaller average frequency of occurrence, although they also present good calibration properties (Table 3).



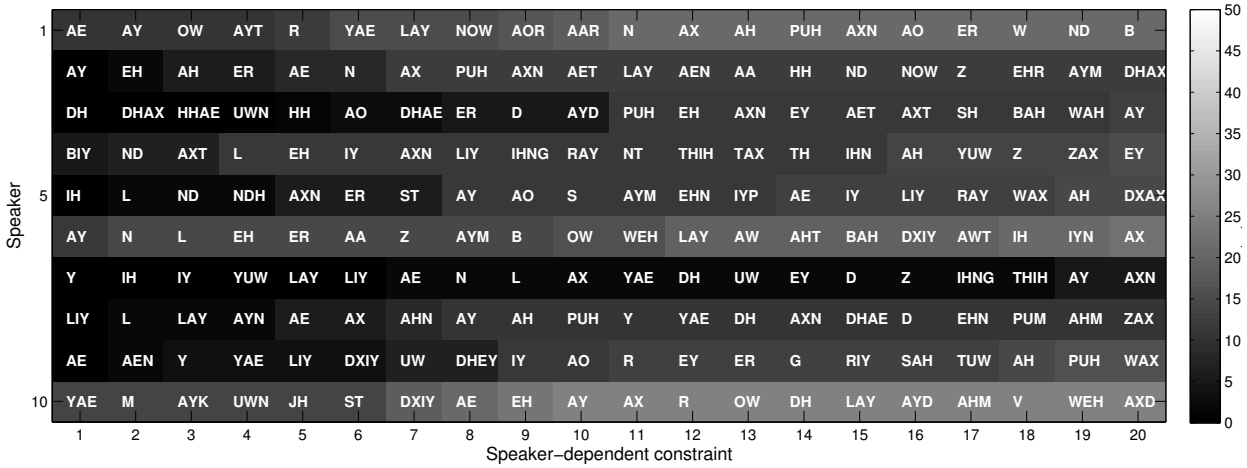
(a) Male trials

(b) Female trials

Figure 7: *EER vs frequency of occurrence for diphone-constraints on the English-only trials of the core condition of the NIST 2006 SRE. Detailed frequency of occurrence in Table B.2.*



(a) Constraints sorted by overall performance



(b) Constraints sorted by speaker-dependent performance (only 10 speakers are shown)

Figure 8: *EER (%) per speaker and constraint (only 20 first constraints are shown) on the English-only male trials of the core condition of the NIST 2006 SRE. In (a), the same unit (columns) performs very differently for different speakers. In (b), for every speaker (rows), the set and order of best constraints vary widely.*

7.1.2. Speaker-dependent performance of different constraints

It is also interesting to analyse how different constraints behave for different speakers, instead of the average behaviour per unit showed in the previous section. While both automatic formant tracking and ASR systems may present different behaviour for different speakers and units, this reflects, in fact, some speaker specificities that are combined with the discriminative abilities of formant frequencies. Figure 8a shows the EER per speaker for the 20-best performing constraints, sorted by overall performance on the evaluation dataset. As the EER has to be computed per each speaker, enough target trials per speaker are needed in order to obtain reliable metrics; with this aim, in this section only those speakers with at least 5 target trials have been used, yielding this 65 male speaker-set (only results for male speakers are shown in Figures 8-11, as similar conclusions can be drawn for female

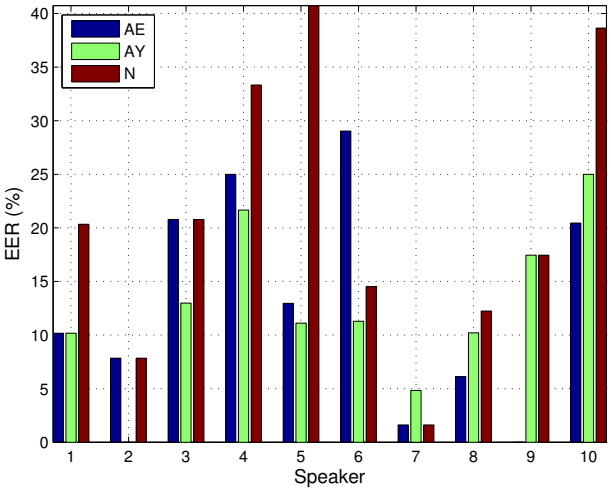


Figure 9: *EER (%) per speaker (only 10 speakers are shown) for the 3-best overall-performing constraints. Missing bars indicates that the EER is equal to zero.*

speakers).

This analysis shows (Figure 8a) that different constraints present different behaviour for different speakers. In fact, the best overall-performing constraint (the phone unit 'AE') may not be the best-performing one for a particular speaker, but even one of the worst-performing. For example, this constraint (first column in Figure 8a) presents a high EER (light grey) for speaker 13 while the performance is much better (dark grey) for speaker 14 and many others. Similarly, the constraint 'AE-T' (last column in Figure 8a), as having a much lower overall-performance (28.42% EER) than the constraint 'AE' (21.21% EER), presents a high EER (light grey) for several speakers, while it still presents a very good performance (dark

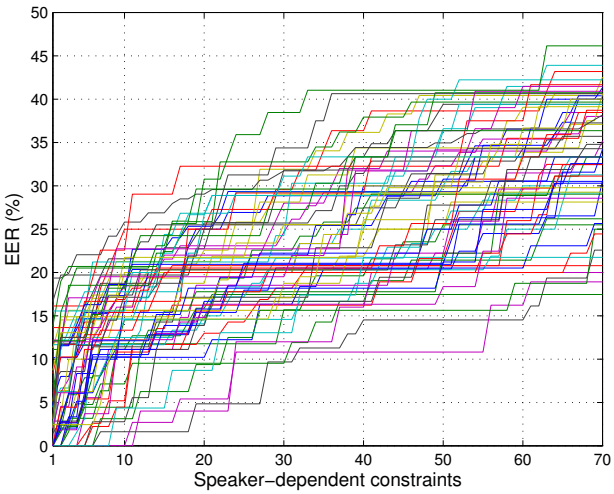


Figure 10: *EER (%) per constraint sorted by speaker-dependent performance for the 65 speakers (each line represents a different speaker). As shown, all 65 speakers have a subset of at least 10 speaker-dependent units with significant discriminative performance (EER per unit below 25%).*

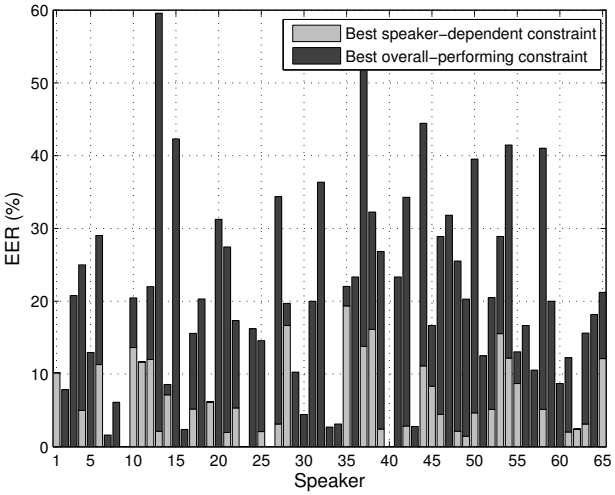


Figure 11: *EER (%) per speaker using the best overall constrained-system and the best speaker-dependent constrained-system. Missing bars indicates that the EER is equal to zero.*

grey) for some others. Similar information is shown in higher detail in Figure 9 in a slightly different way. Here, the EER of the 3-best overall-performing constraints ('AE', 'AY' and 'N') is represented for 10 different speakers, showing their highly variable performance from one speaker to another.

Moreover, when constraints are sorted by performance independently for each speaker, the set of N best-performing constraints can be very different from one speaker to another, as it can be seen in Figure 8b (where only 10 speakers and their 20 best-performing constraints are shown). This analysis also shows that a very good performance (low EER) could be achieved for most of the speakers if a speaker-dependent set of constraints is used, as it is shown in Figure 10 for all 65 speakers and their 70 best-performing constraints (conversely to Figure 8b, the particular constraints are not shown). It can be seen that, for every speaker, there is at least one constraint (and usually between 5 and 10 constraints) with better performance than the best overall-performing constraint (21.21% EER). Moreover, all 65 speakers have a subset of at least 10 speaker-dependent units with significant discriminative performance (EER per unit below 25%).

As an independent system is built for each isolated constraint in this approach, it would be possible to take advantage of this fact by using a different linguistically-constrained system for each speaker in order to adapt to his/her particular specificities if they were known in advance. For example, in the NIST 2012 SRE the target speakers were known in advance and several utterances per target speaker were provided; similar conditions may exist in real-life applications like access control or wiretapping. In such a case, the performance of the different constrained-systems could be analysed for each target speaker on a development dataset.

Figure 11 shows how the EER per speaker could be highly improved if the best constraint is selected in a speaker-dependent way instead of taking the best overall-performing constraint. While for these 65 speakers the average EER using the best overall-performing

constraint ('AE' phone) is 19.49%, the average EER using the best-performing constraint of each speaker would be 4.10%, a remarkable result as, for this speaker-set, the average EER of the reference cepstral system is 3.31%. Although this last result is optimistic as it is obtained knowing the best-performing speaker-dependent units over the evaluation dataset, it shows that improved results could be obtained adopting speaker-dependent strategies.

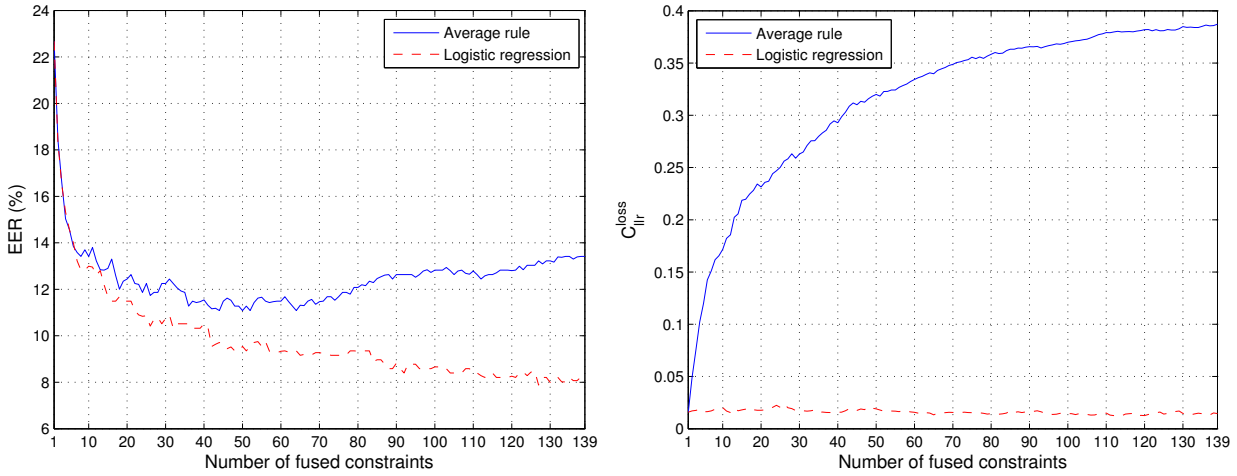
7.2. Performance of speaker-independent combinations of constraints

7.2.1. Comparison of fusion techniques

Figure 12a shows the EER of the fused system as a function of the number of fused constraints on the sre05-val development dataset for male trials for the two fusion techniques analysed in this work (namely, the average rule and logistic regression). While the EER of the fused system through the average rule obtains a minimum value for a certain number of fused constraints and then begin to increase, the EER of the fused system through logistic regression keeps going down as the number of fused constraints increases. The logistic regression fusion, being a trained fusion rule, benefits from the increasing amount of data provided by the additional constraints to be fused.

However, these are optimistic results as they are obtained in the development dataset, and the combination of constraints on the evaluation dataset can degrade if the performance of fused constraints varies from that obtained in development. This effect can be seen in Table 4. For the logistic regression technique, while the EER of the best fused system on the development dataset decreases as long as we take into account more constraints (from 41 phones to 41 phones + 98 diphones), the difference with the evaluation results increases, making them less robust to dataset mismatch for a large number of fused constraints. Conversely, the average fusion rule benefits from a higher number of constraints even in the case of dataset mismatch.

On the other hand, it can be seen also from Figure 12b that the calibration loss increases



(a) EER of the fused system

(b) C_{llr}^{loss} of the fused system

Figure 12: *Comparison of fusion techniques on male trials of the sre05-val development dataset.*

		Male/female EER (%) for the **N-best** rule		
		Phones (P)	Diphones (D)	P+D
Average	sre05-val	12.35 / 13.67	12.63 / 15.06	11.08 / 13.55
	SRE06	10.33 / 14.82	12.72 / 15.10	9.93 / 13.50
Log. Reg.	sre05-val	11.68 / 12.14	10.51 / 12.23	7.88 / 9.84
	SRE06	**9.57** / 12.89	12.72 / 15.36	11.26 / **12.62**

Table 4: *EER (%) for male/female trials in development and evaluation datasets when combining different types of linguistic units through the N-best rule.*

		Male/female EER (%) for the **SFS** rule		
		Phones (P)	Diphones (D)	P+D
Average	sre05-val	12.25 / 13.18	11.87 / 14.56	10.70 / 12.58
	SRE06	10.17 / 14.45	12.99 / 16.15	**9.66** / **13.89**
Log. Reg.	sre05-val	11.87 / 12.58	11.39 / 13.91	10.70 / 12.08
	SRE06	11.15 / 14.11	12.34 / 15.72	10.33 / 14.17

Table 5: *EER (%) for male/female trials in development and evaluation datasets when combining different types of linguistic units through the SFS rule.*

for the average fusion as the number of fused constraint increases, while it remains almost constant for the logistic regression. This makes the logistic regression the preferred fusion option as eliciting calibrated LRs is among our main objectives.

7.2.2. Comparison of constraint-selection strategies

Table 5 shows the results for the SFS constraint-selection strategy as Table 4 does for the N-best one. It can be seen that both strategies give similar results on the evaluation dataset for the average fusion rule, being the EER of the fused systems reduced when constraints from different linguistic-unit types (phones and diphones) are combined. However, in the case of the logistic regression fusion technique, there is no such gain for the N-best strategy on male trials and slight differences on female trials due to the over-fitting and database mismatch between development and evaluation datasets observed in the previous section. The SFS strategy does not suffer from this over-fitting as it does not select a number of constraints as high as the N-best strategy, as constraints that do not increase the performance of the fused system are discarded. In this way, it still benefits from incorporating diphone units, which can provide additional dynamic information present in the transition between phone units.

Finally, Table 6 shows the performance on different evaluation metrics for the best combinations of constraint-selection strategies and fusion techniques. In this table, we can see that logistic regression technique has the advantage of providing well calibrated likelihood

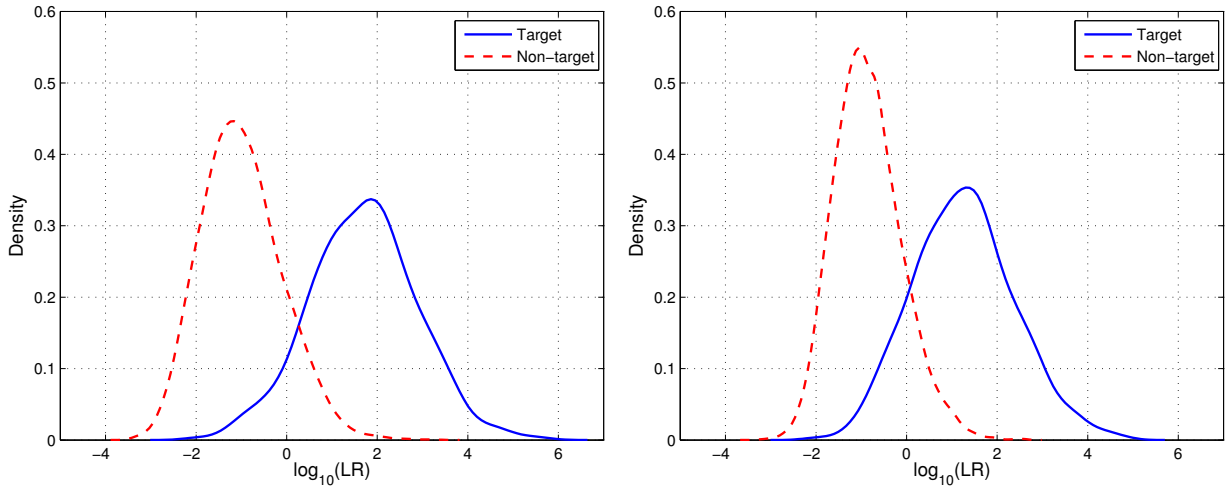
		Male/female results on the **NIST 2006 SRE, English-only trials**			
		EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
SFS	Phones	10.17 / 14.45	0.0495 / 0.0585	0.6759 / 0.6928	0.3069 / 0.2229
Average	P+D	9.66 / 13.89	0.0463 / 0.0585	0.8163 / 0.8439	0.4741 / 0.3975
N-best	Phones	**9.57** / 12.89	**0.0456** / **0.0543**	**0.3742** / **0.4361**	**0.0277** / **0.0117**
Log. reg.	P+D	11.26 / **12.62**	0.0503 / 0.0590	0.4046 / 0.4531	0.0317 / 0.0202

Table 6: *Comparison of the best combinations between constraint-selection strategies and fusion techniques on the evaluation dataset.*

ratios also on the evaluation dataset, as we saw in Figure 12 for the development dataset. Being this a highly desirable property, the following analysis in Section 7.3 focus on the best combination of constraints through logistic regression, which is the one using N-best selection from phone-constraints. In order to highlight the calibration properties of the elicited LRs from the best formant-based fused system, in Figure 13 we show the $\log_{10}$LR target and non-target probability density functions.

7.3. Fusion of formant- and cepstral-based systems

Table 7 show the results on the evaluation dataset for the best formant-based fused system (that using logistic regression fusion of the N-best selected phone units), for the cepstral-based reference system, and for the average fusion of both. For female trials, although the EER of the fused system is almost the same, there are significant improvements



(a) Male trials (b) Female trials

Figure 13: *Target and non-target $\log_{10}$-LRs probability density functions for the best formant-based fused system on the English-only trials of the core condition of the NIST 2006 SRE.*

	Male/female results on the **NIST 2006 SRE, English-only trials**			
	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
Formant-based	9.57 / 12.89	0.0456 / 0.0543	0.3742 / 0.4361	0.0277 / **0.0117**
Cepstral-based	6.21 / 6.87	0.0303 / 0.0352	0.2293 / 0.2927	**0.0232** / 0.0321
Average fusion	**5.41 / 6.86**	**0.0248 / 0.0311**	**0.2179 / 0.2789**	0.0368 / 0.0393

Table 7: *Results on the evaluation dataset for the best formant-based fused system, the cepstral-based reference system and the average fusion of both.*

in both minDCF and C_{llr} metrics. In the case of male trials, there is also a relative improvement of ∼18% in terms of the EER. Considering both genders, the fused system obtains relative improvements of 7% and 15% in terms of gender-averaged EER and minDCF, respectively. Although both approaches are based on spectral features, it is shown that they present a high complementarity like other high-level approaches on the same evaluation framework [49] [12] [15].

7.4. Comparison with other higher-level systems

Finally, an objective comparison of different high-level approaches in the same evaluation framework (core condition of the NIST 2006 SRE) is given in Table 8, extending that presented in [10] with some later works, sorted by performance.

The best performing systems are those based on cepstral information (1,2), using either cepstral-derived features (coefficients from MLLR transforms between cepstral-based GMMs) or MFCC (and prosodic) contours, where ASR is used either only for feature extraction (2) or also for region conditioning (1). Then, there is a group of systems based on several prosodic (usually including energy, pitch and duration) and/or formant features (3-7), most of them having very similar performance (ranging from 10.41% to 11.9% EER for the four best performing ones). Next two systems are based only on duration information: (8) models the number of frames of the three states in phone HMMs, while (9) directly models the duration of phones within specific words. Finally, system (10) is a lexico-prosodic approach with similar performance to (9).

Among them, our approach is the only one based only on formant frequencies and where feature extraction does not rely on ASR labels, which are only used for region conditioning. Also, it is worth noting that our formant-based system does not include NIST 2005 SRE in the background dataset in order to avoid using overoptimistic scores in the calibration training; in this way, it is possible to obtain well calibrated LRs per constraint, but better discriminative performance may be achieved using a richer and larger background database for UBM and total variability training. However, being its features obtained from short-term windows every 10 ms, system parameters can be properly trained on limited background data.

Male+female results on the **NIST 2006 SRE, English-only trials**		
System (feature type and model)	EER (%)	Reference
(1) Cepstral-derived MLLR SVM	4.00	[51]
(2) Prosodic and MFCC contours JFA	7.66	[15]
(3) Syllable-based prosody sequence SVM	10.41	[52], [53]
(4) Prosodic contours JFA	11.00	[13]
(5) **Formants+Δs i-vector**	**11.23**	-
(6) Formant and prosodic contours JFA	11.9	[12]
(7) Prosodic contours JFA	14.6	[49]
(8) State-in-phone-duration GMM	16.02	[54]
(9) Phone-in-word-duration GMM	22.22	[54]
(10) Duration-conditioned word N-gram SVM	23.46	[55]

Table 8: *Results on the core condition of the NIST 2006 SRE (English-only trials) for several high-level systems compared to our formant-based approach (5).*

8. Conclusions and future work

In this work, we have explored the discriminative abilities of formant frequencies and their dynamics within linguistic units through fully-automatic linguistically-constrained i-vector systems.

Automatic formant tracking have been used for feature extraction, and dynamic information is included through derivative coefficients. In this way, it is possible to combine both static and dynamic information of formant frequencies while maintaining the frame-by-frame feature observation rate, instead of reducing each constraint to a single observation feature vector as it is done in some approaches that code the whole trajectory within a speech region. This procedure allows us to robustly train the parameters of the system even with limited background data (NIST SRE 2004) compared with similar higher-level approaches based on coded trajectories, as it has been shown in Section 7.4.

Then, ASR is used in order to constrain the set of features to be used by each subsystem, corresponding each of them to a different linguistic unit among two main groups: phones and diphones. For each of such constraints, one independent i-vector system is developed. Although linguistically-constrained systems have limited performance by themselves, we have shown that well calibrated log-likelihood ratios can be provided for each linguistic unit. Regarding the relative differences in performance among units, it should be noted that they can be due not only to the different discriminative abilities of their formant frequencies but also to a different behaviour of the automatic systems involved in the feature extraction (formant tracking) and region conditioning (ASR labels) processes, which may lead to a non-uniform distribution of errors among different units. It would be of broad interest to perform an equivalent analysis in a manually labelled database in order to avoid the effect of the errors introduced by these automatic systems, but large datasets of spontaneous conversational speech as those used in this work ($\sim$10,000 5-minute conversations) seem unlikely to be manually annotated (both formant frequencies and phonetic transcriptions).

On the other hand, a different behaviour of the formant tracking and ASR systems across speakers for a particular unit is considered to reflect the specificities of the different speakers.

This fine-grained detail provided by linguistically-constrained systems can be exploited through speaker-dependent strategies when selecting the constraints to be used. For example, in Section 7.1.2 it has been shown that using only the best-performing speaker-dependent constraint instead of the best overall-performing one for every speaker, the average EER in the analysed speaker set improves from 19.49% to 4.10%. Furthermore, most of the speakers in the analysed set presents a subset of several constraints (usually between 5 and 10) that perform better than the overall-performing constraint, so using any of those (different) constraints for every speaker will lead to an overall performance improvement. Although this is an optimistic result as it is obtained knowing the best-performing speaker-dependent units over the evaluation dataset, it shows that improved results could be obtained adopting speaker-dependent strategies. As a future work, some of this strategies would be tested on an experimental framework that allows to estimate in advance the best speaker-dependent set of linguistic units to be used for the different target speakers.

Moreover, we have presented several speaker-independent constraint-combination strategies in order to integrate the speaker distinguishing information spread over the different linguistic units, achieving for some of them a remarkable combined performance taking into account the limited size of the background dataset and the nature of features being used. For these fused systems, discriminative and well calibrated log-likelihood ratios are also provided.

Finally, significant improvements have been achieved by combining these formant-based systems with a cepstral-based reference system, showing the complementarity of cepstral and formant-based approaches.

9. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz, TEC2012-37585-C02-01). Also, the authors would like to thank SRI for providing the Decipher phonetic transcriptions of the NIST 2004, 2005 and 2006 SREs that have allowed to carry out this work.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] J. Darch, B. Milner, X. Shao, S. Vaseghi, and Q. Yan, "Predicting formant frequencies from MFCC vectors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 2005, pp. 941–944.

[3] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law*, vol. 12, no. 2, 2005.

[4] P. Rose, *Forensic Speaker Identification*, ser. Forensic Science. Taylor and Francis, 2002.

[5] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and gaussian mixture models," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia, September 22-26, 2008*, 2008, pp. 1505–1508.

[6] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. 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. 2104–2115, Sept 2007.

[7] F. Nolan, *The phonetic bases of speaker recognition.* Cambridge (UK): Cambridge University Press, 1983.

[8] K. McDougall, "Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies," *International Journal of Speech Language and the Law*, vol. 13, no. 1, pp. 89 – 126, 2006.

[9] J. Gonzalez-Rodriguez, J. Gil, R. Pérez, and J. Franco-Pedroso, "What are we missing with i-vectors? a perceptual analysis of i-vector-based falsely accepted trials," in *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 33–40.

[10] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I. Fundamentals, Features, and Methods*, ser. Lecture Notes in Computer Science, vol. 4343. Springer Berlin Heidelberg, 2007, pp. 241–259.

[11] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4525–4528.

[12] N. Dehak, P. Kenny, and P. Dumouchel, "Continuous prosodic features and formant modeling with joint factor analysis for speaker verification," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 1234–1237.

[13] M. Kockmann, L. Burget, and J. Cernocký, "Investigations into prosodic syllable contour features for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, 2010, pp. 4418–4421.

[14] D. A. Reynolds, W. D. Andrews, J. P. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. S. Abramson, R. Mihaescu, J. J. Godfrey, D. A. Jones, and B. Xiang, "The supersid project: exploiting high-level information for high-accuracy speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, April 6-10, 2003*, 2003, pp. 784–787.

[15] M. Kockmann and L. Burget, "Contour modeling of prosodic and acoustic features for speaker recognition," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT 2008)*, Dec 2008, pp. 45–48.

[16] J. Gonzalez-Rodriguez, "Speaker recognition using temporal contours in linguistic units: The case of formant and formant-bandwidth trajectories," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy, August 27-31, 2011*, 2011, pp. 133–136.

[17] J. Franco-Pedroso, F. Espinoza-Cuadros, and J. Gonzalez-Rodriguez, "Formant trajectories in linguistic units for text-independent speaker recognition," in *Proceedings of the International Conference on Biometrics (ICB 2013), 4-7 June, 2013, Madrid, Spain*, 2013, pp. 1–6.

[18] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[19] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 36, no. 6, pp. 871–879, Jun 1988.

[20] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing.* Berlin:

Springer, 2008.
[21] P. Rose and E. Winter, "Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology (SST 2010)*, December 2010, pp. 42–45.
[22] C. Zhang, G. S. Morrison, and P. Rose, "Forensic speaker recognition in chinese: a multivariate likelihood ratio discrimination on /i/ and /y/," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia, September 22-26, 2008*, 2008, pp. 1937–1940.
[23] F. Nolan, "The 'telephone effect' on formants: a response," *International Journal of Speech Language and the Law*, vol. 9, no. 1, 2002.
[24] G. S. Morrison, "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories," *Journal of the Acoustical Society of America*, no. 125, pp. 2387–2397, 2009.
[25] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000 / INTERSPEECH), Beijing, China, October 16-20, 2000*, 2000, pp. 464–467.
[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: http://www.praat.org/
[27] Y. Laprie, "Winsnoori 1.34 – free software for speech analysis," 2014. [Online]. Available: http://www.loria.fr/~laprie/WinSnoori/
[28] "Snack Sound Toolkit — Wikipedia, The Free Encyclopedia," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Snack_Sound_Toolkit
[29] "Tcl — Wikipedia, The Free Encyclopedia," 2015. [Online]. Available: http://en.wikipedia.org/wiki/Tcl
[30] "Snack v2.2.8 manual." [Online]. Available: http://www.speech.kth.se/snack/man/snack2.2/tcl-man.html
[31] "Arpabet — Wikipedia, The Free Encyclopedia," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Arpabet
[32] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Proceedings of the IEEE Workshop on Signal Processing Applications for Public Security and Forensics (SAFE 2007)*, April 2007, pp. 1–5.
[33] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.
[34] S. S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4205–4208.
[35] L. Rabiner and B. Juang, "Speech recognition: Statistical methods," in *Encyclopedia of Language and Linguistics*, K. Brown, Ed. Elsevier, 2006.
[36] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.
[37] J. E. Shoup, "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, W. A. Lea, Ed. Englewood Cliffs: Prentice Hall, 1980, pp. 125–138.
[38] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252.
[39] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.
[40] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," in *Computer*

Speech and Language, vol. 20, no. 2 - 3, 2006, pp. 230 – 275.

[41] N. Brümmer, "Toolkit for evaluation, fusion and calibration of statistical pattern recognizers." [Online]. Available: https://sites.google.com/site/nikobrummer/focal

[42] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, 2000.

[43] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.

[44] N. Brümmer, L. Burget, J. Cernocký, O. Glembek, F. Grézl, M. Karafiát, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[45] J. Gonzalez-Rodriguez, "Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014)," *Loquens*, vol. 1, no. 1, pp. 1–15, January 2014.

[46] "The NIST Year 2004 Speaker Recognition Evaluation Plan." [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2004/SRE-04_evalplan-v1a.pdf

[47] "The NIST Year 2005 Speaker Recognition Evaluation Plan." [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2005/sre-05_evalplan-v6.pdf

[48] "The NIST Year 2006 Speaker Recognition Evaluation Plan." [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2006/sre-06_evalplan-v9.pdf

[49] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, Sept 2007.

[50] D. van Leeuwen and N. Brümmer, "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems," in *Speaker Classification I*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Springer Berlin Heidelberg, 2007, vol. 4343, pp. 330–353.

[51] A. Stolcke, L. Ferrer, S. S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005 - EUROSPEECH), Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 2425–2428.

[52] L. Ferrer, E. Shriberg, S. S. Kajarekar, and M. K. Sönmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Honolulu, Hawaii, USA, April 15-20, 2007*, 2007, pp. 233–236.

[53] E. Shriberg and L. Ferrer, "A text-constrained prosodic system for speaker verification," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 1226–1229.

[54] L. Ferrer, H. Bratt, V. R. R. Gadde, S. S. Kajarekar, E. Shriberg, M. K. Sönmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH 2003 - EUROSPEECH), Geneva, Switzerland, September 1-4, 2003*, 2003.

[55] G. Tür, E. Shriberg, A. Stolcke, and S. S. Kajarekar, "Duration and pronunciation conditioned lexical modeling for speaker verification," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 2049–2052.

Appendix A. Extended results

Table A.1: *Results on the evaluation dataset for every phone-constraint.*

NIST 2006 SRE, English-only trials									
Male					Female				
Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AA	31.21	0.0971	0.8397	0.0112	AA	32.20	0.0982	0.8707	0.0113
AE	21.21	0.0850	0.6668	0.0143	AE	24.59	0.0876	0.7308	0.0131
AH	23.96	0.0964	0.7286	0.0158	AH	28.89	0.0974	0.8185	0.0079
AO	28.47	0.0996	0.8264	0.0151	AO	33.08	0.0989	0.8767	0.0115
AW	33.88	0.0986	0.8907	0.0108	AW	37.48	0.1000	0.9352	0.0075
AX	23.80	0.0844	0.7001	0.0150	AX	27.15	0.0932	0.7764	0.0100
AY	21.38	0.0825	0.6580	0.0158	AY	24.59	0.0841	0.7101	0.0111
B	33.74	0.0956	0.8774	0.0141	B	34.28	0.0988	0.8851	0.0090
CH	41.57	0.0997	0.9665	0.0102	CH	41.87	0.0999	0.9757	0.0061
D	31.21	0.0951	0.8333	0.0130	D	34.28	0.0982	0.8868	0.0097
DH	26.16	0.0932	0.7514	0.0125	DH	27.79	0.0940	0.7876	0.0114
DX	37.59	0.0997	0.9289	0.0134	DX	39.35	0.1000	0.9613	0.0093
EH	24.83	0.0972	0.7544	0.0140	EH	28.06	0.0990	0.8196	0.0157
ER	29.40	0.0996	0.8409	0.0157	ER	33.53	0.0996	0.8816	0.0144
EY	28.42	0.0995	0.8164	0.0132	EY	31.96	0.0978	0.8608	0.0080
F	37.47	0.0986	0.9288	0.0095	F	42.22	0.0995	0.9701	0.0109
G	35.45	0.1000	0.9126	0.0137	G	38.97	0.0999	0.9416	0.0078
HH	31.37	0.0980	0.8624	0.0149	HH	34.46	0.0980	0.8933	0.0112
IH	26.48	0.0931	0.7736	0.0174	IH	31.18	0.0998	0.8554	0.0160
IY	27.25	0.0996	0.7854	0.0151	IY	30.44	0.0975	0.8384	0.0128
JH	37.96	0.0994	0.9338	0.0151	JH	38.58	0.1000	0.9457	0.0098
K	34.13	0.0979	0.8779	0.0101	K	37.51	0.0994	0.9263	0.0109
L	23.24	0.0839	0.7083	0.0133	L	24.68	0.0869	0.7355	0.0127
M	31.21	0.0961	0.8445	0.0114	M	28.89	0.0944	0.8169	0.0121
N	22.26	0.0812	0.6896	0.0168	N	24.77	0.0839	0.7256	0.0112
NG	36.40	0.0975	0.9139	0.0165	NG	33.77	0.0991	0.8809	0.0110
OW	25.49	0.0927	0.7445	0.0152	OW	27.79	0.0936	0.7830	0.0098
P	39.99	0.0998	0.9439	0.0082	P	41.96	0.0999	0.9732	0.0069
PUH	24.32	0.0933	0.7296	0.0137	PUH	31.28	0.0976	0.8420	0.0097
PUM	34.86	0.0983	0.8985	0.0217	PUM	32.46	0.0984	0.8764	0.0142
R	24.96	0.0937	0.7380	0.0149	R	26.49	0.0932	0.7681	0.0132
S	31.48	0.0947	0.8390	0.0095	S	34.18	0.0963	0.8840	0.0123
SH	36.73	0.1000	0.9255	0.0121	SH	40.40	0.0998	0.9571	0.0122
T	29.98	0.0925	0.8247	0.0161	T	32.37	0.0959	0.8624	0.0095
TH	37.10	0.0978	0.9387	0.0159	TH	39.31	0.1000	0.9506	0.0054
UH	39.64	0.0999	0.9471	0.0075	UH	40.31	0.1000	0.9593	0.0106
UW	28.02	0.0950	0.8016	0.0121	UW	33.63	0.0993	0.8980	0.0121
V	37.39	0.0998	0.9286	0.0106	V	38.49	0.0999	0.9456	0.0082
W	31.21	0.0948	0.8307	0.0111	W	30.81	0.0973	0.8509	0.0177
Y	24.68	0.0915	0.7325	0.0180	Y	30.08	0.0923	0.8303	0.0124
Z	32.27	0.0952	0.8658	0.0150	Z	35.20	0.0997	0.9083	0.0091

Table A.2: *Results on the evaluation dataset for every diphone-constraint.*

NIST 2006 SRE, English-only trials									
Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AA-R	37.99	0.1000	0.9414	0.0085	AA-R	38.03	0.0988	0.9355	0.0123
AA-T	34.80	0.0982	0.9027	0.0175	AA-T	34.46	0.1000	0.9020	0.0122
AE-N	28.88	0.0945	0.8132	0.0129	AE-N	31.71	0.0963	0.8529	0.0098
AE-T	28.42	0.0987	0.8103	0.0195	AE-T	31.90	0.0997	0.8614	0.0128
AE-V	38.53	0.1000	0.9433	0.0131	AE-V	39.77	0.0998	0.9554	0.0078
AH-M	36.93	0.0996	0.9249	0.0120	AH-M	39.38	0.1000	0.9505	0.0090
AH-N	36.67	0.0996	0.9320	0.0102	AH-N	38.10	0.1000	0.9419	0.0124
AH-T	35.20	0.0986	0.9040	0.0137	AH-T	36.02	0.0996	0.9101	0.0102
AO-L	36.73	0.1000	0.9312	0.0101	AO-L	39.22	0.1000	0.9530	0.0092
AO-R	34.60	0.0992	0.8978	0.0149	AO-R	33.28	0.0990	0.8909	0.0103
AW-T	39.32	0.0980	0.9406	0.0190	AW-T	40.59	0.0999	0.9600	0.0095
AX-B	40.39	0.0998	0.9593	0.0118	AX-B	42.38	0.1000	0.9727	0.0102
AX-D	39.94	0.1000	0.9668	0.0105	AX-D	41.50	0.0999	0.9695	0.0070
AX-G	40.57	0.1000	0.9682	0.0088	AX-G	44.77	0.0997	0.9825	0.0078
AX-K	38.26	0.0991	0.9391	0.0102	AX-K	42.51	0.1000	0.9678	0.0077
AX-L	37.07	0.0985	0.9327	0.0123	AX-L	37.00	0.1000	0.9297	0.0083
AX-M	43.84	0.0999	0.9930	0.0204	AX-M	41.71	0.0999	0.9724	0.0068
AX-N	27.79	0.0947	0.7829	0.0132	AX-N	29.34	0.0942	0.8190	0.0154
AX-NG	35.18	0.1000	0.9294	0.0154	AX-NG	38.12	0.1000	0.9321	0.0080
AX-S	34.53	0.0983	0.8957	0.0149	AX-S	36.77	0.0999	0.9240	0.0098
AX-T	33.26	0.1000	0.8946	0.0126	AX-T	38.37	0.0999	0.9482	0.0111
AX-V	43.84	0.1000	0.9835	0.0084	AX-V	44.84	0.0999	0.9873	0.0049
AX-Z	37.86	0.1000	0.9430	0.0113	AX-Z	38.85	0.1000	0.9523	0.0094
AY-D	36.13	0.0987	0.9185	0.0157	AY-D	40.59	0.0996	0.9486	0.0074
AY-K	32.27	0.0976	0.8699	0.0153	AY-K	33.65	0.0988	0.8825	0.0119
AY-M	32.67	0.0970	0.8616	0.0122	AY-M	35.01	0.0993	0.9157	0.0161
AY-N	35.37	0.0988	0.9096	0.0139	AY-N	39.50	0.1000	0.9493	0.0078
AY-T	35.07	0.0989	0.9006	0.0154	AY-T	35.09	0.0998	0.9146	0.0129
B-AH	37.12	0.0993	0.9309	0.0148	B-AH	35.90	0.0998	0.9210	0.0141
B-AX	39.46	0.0998	0.9588	0.0146	B-AX	41.59	0.1000	0.9682	0.0085
B-IY	38.95	0.0997	0.9476	0.0141	B-IY	39.40	0.1000	0.9444	0.0081
D-AX	39.99	0.1000	0.9585	0.0071	D-AX	38.75	0.1000	0.9458	0.0076
D-DH	41.49	0.0998	0.9670	0.0103	D-DH	40.75	0.1000	0.9574	0.0098
DH-AE	28.82	0.0950	0.8270	0.0172	DH-AE	33.47	0.0954	0.8769	0.0127
DH-AX	33.07	0.0996	0.8868	0.0123	DH-AX	35.11	0.1000	0.9142	0.0069
DH-EH	36.73	0.0998	0.9281	0.0134	DH-EH	36.29	0.0999	0.9174	0.0124
DH-EY	35.07	0.0996	0.9194	0.0121	DH-EY	37.12	0.0990	0.9306	0.0111
D-IH	40.95	0.0999	0.9665	0.0119	D-IH	40.74	0.1000	0.9614	0.0076
D-OW	40.47	0.0998	0.9644	0.0099	D-OW	40.81	0.0990	0.9621	0.0161
D-UW	40.27	0.1000	0.9573	0.0133	D-UW	41.62	0.1000	0.9718	0.0097
DX-AX	43.52	0.0998	0.9782	0.0096	DX-AX	42.51	0.1000	0.9782	0.0078
DX-IY	39.19	0.0995	0.9462	0.0112	DX-IY	41.36	0.1000	0.9742	0.0086
EH-L	37.47	0.0998	0.9409	0.0139	EH-L	37.40	0.0999	0.9453	0.0109
EH-N	33.34	0.0999	0.8867	0.0150	EH-N	34.65	0.1000	0.8986	0.0120
Continued on next page									

Table A.2 – continued from previous page

Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
EH-R	32.27	0.0998	0.8820	0.0111	EH-R	36.53	0.0997	0.9183	0.0125
HH-AE	33.39	0.0993	0.8873	0.0162	HH-AE	34.19	0.1000	0.8944	0.0124
HH-W	44.68	0.1000	0.9875	0.0061	HH-W	44.70	0.1000	0.9878	0.0089
IH-N	37.72	0.0996	0.9299	0.0105	IH-N	37.39	0.1000	0.9261	0.0100
IH-NG	33.47	0.0998	0.8902	0.0156	IH-NG	33.65	0.0995	0.8833	0.0107
IH-T	35.55	0.0996	0.9216	0.0162	IH-T	37.12	0.0998	0.9347	0.0095
IY-AX	38.49	0.0992	0.9368	0.0117	IY-AX	43.52	0.1000	0.9810	0.0075
IY-N	37.47	0.0969	0.9225	0.0126	IY-N	41.31	0.1000	0.9727	0.0060
IY-P	41.98	0.1000	0.9794	0.0137	IY-P	43.52	0.0999	0.9768	0.0093
JH-AX	39.85	0.0994	0.9556	0.0169	JH-AX	39.86	0.1000	0.9574	0.0144
K-AH	39.99	0.0997	0.9510	0.0108	K-AH	40.40	0.1000	0.9566	0.0115
K-AX	39.81	0.0997	0.9573	0.0083	K-AX	41.65	0.1000	0.9738	0.0056
K-S	44.51	0.1000	0.9915	0.0069	K-S	47.25	0.1000	0.9948	0.0054
L-AX	41.98	0.1000	0.9690	0.0122	L-AX	40.96	0.1000	0.9630	0.0061
L-AY	28.42	0.0957	0.8216	0.0277	L-AY	29.71	0.0944	0.8154	0.0199
L-IY	30.15	0.0966	0.8331	0.0130	L-IY	32.20	0.1000	0.8728	0.0114
M-AX	43.58	0.0998	0.9829	0.0103	M-AX	44.98	0.1000	0.9850	0.0076
M-AY	42.25	0.1000	0.9683	0.0110	M-AY	41.05	0.0988	0.9574	0.0110
M-IY	38.12	0.0984	0.9547	0.0248	M-IY	38.12	0.0997	0.9365	0.0122
N-AA	35.86	0.0972	0.9145	0.0157	N-AA	39.22	0.1000	0.9497	0.0088
N-AX	38.21	0.1000	0.9409	0.0136	N-AX	39.86	0.1000	0.9556	0.0066
N-D	31.80	0.0974	0.8469	0.0152	N-D	33.00	0.0978	0.8736	0.0082
N-DH	39.91	0.0998	0.9534	0.0113	N-DH	40.56	0.1000	0.9615	0.0079
NG-K	40.12	0.1000	0.9598	0.0103	NG-K	42.18	0.1000	0.9658	0.0086
N-IY	41.86	0.0993	0.9752	0.0189	N-IY	37.65	0.1000	0.9395	0.0087
N-OW	32.15	0.0959	0.8665	0.0128	N-OW	30.71	0.0958	0.8490	0.0159
N-S	42.82	0.0999	0.9798	0.0111	N-S	43.40	0.1000	0.9809	0.0057
N-T	35.35	0.0990	0.9099	0.0147	N-T	36.57	0.0997	0.9178	0.0096
OW-N	40.39	0.0993	0.9540	0.0116	OW-N	41.05	0.1000	0.9608	0.0092
P-AX	42.25	0.0997	0.9726	0.0107	P-AX	42.15	0.0998	0.9722	0.0093
R-AX	42.69	0.0999	0.9804	0.0163	R-AX	42.60	0.1000	0.9760	0.0053
R-AY	34.88	0.0996	0.9107	0.0239	R-AY	33.65	0.0999	0.8984	0.0174
R-IY	36.25	0.1000	0.9160	0.0138	R-IY	35.20	0.0999	0.9113	0.0124
S-AH	37.99	0.0994	0.9306	0.0124	S-AH	39.95	0.0999	0.9468	0.0099
S-AX	40.92	0.0997	0.9546	0.0114	S-AX	39.00	0.1000	0.9452	0.0083
S-OW	36.13	0.0992	0.9107	0.0102	S-OW	33.22	0.0996	0.8811	0.0116
S-T	38.92	0.0981	0.9333	0.0107	S-T	38.75	0.0999	0.9527	0.0105
T-AX	37.08	0.0985	0.9273	0.0123	T-AX	35.43	0.0999	0.9095	0.0096
T-AY	38.20	0.0983	0.9393	0.0116	T-AY	37.50	0.0996	0.9288	0.0138
T-DH	41.06	0.0995	0.9583	0.0127	T-DH	40.56	0.1000	0.9597	0.0056
TH-IH	36.80	0.0986	0.9122	0.0253	TH-IH	36.11	0.0999	0.9187	0.0106
T-R	43.70	0.0999	0.9832	0.0055	T-R	41.39	0.1000	0.9685	0.0054
T-S	41.06	0.0995	0.9618	0.0081	T-S	40.54	0.0998	0.9591	0.0086
T-UW	38.12	0.0997	0.9393	0.0127	T-UW	37.12	0.0995	0.9279	0.0123
T-W	44.38	0.1000	0.9875	0.0099	T-W	43.33	0.1000	0.9869	0.0073
UH-D	42.45	0.1000	0.9712	0.0090	UH-D	41.02	0.1000	0.9678	0.0067

Continued on next page

Table A.2 – continued from previous page

Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
UW-N	34.15	0.0935	0.8771	0.0161	UW-N	33.46	0.0967	0.8781	0.0140
V-AX	40.79	0.1000	0.9763	0.0100	V-AX	42.33	0.1000	0.9718	0.0074
W-AH	37.70	0.1000	0.9415	0.0095	W-AH	37.66	0.1000	0.9341	0.0092
W-AX	41.31	0.1000	0.9632	0.0115	W-AX	44.46	0.1000	0.9831	0.0090
W-EH	38.78	0.0992	0.9397	0.0101	W-EH	40.72	0.0999	0.9614	0.0083
Y-AE	25.26	0.0867	0.7533	0.0245	Y-AE	28.32	0.0894	0.7960	0.0130
Y-UW	27.69	0.0928	0.8005	0.0135	Y-UW	32.65	0.0948	0.8667	0.0114
Z-AX	37.99	0.0986	0.9306	0.0111	Z-AX	39.07	0.1000	0.9478	0.0070

Appendix B. Detailed frequency of occurrence

Table B.1: *Frequency of occurrence in NIST SRE 2004, 2005 and 2006 conversations for every phone-constraint.*

Average number of occurrences per conversation								
Phone	Male	Female	Phone	Male	Female	Phone	Male	Female
AA	17.8	16.2	EY	23.5	22.3	PUH	17.9	13.8
AE	45.1	41.4	F	19.8	17.6	PUM	6.3	8.6
AH	33.2	31.0	G	13.3	12.5	R	47.4	43.9
AO	17.7	15.9	HH	20.3	22.7	S	61.1	56.7
AW	8.1	7.6	IH	42.4	39.6	SH	7.1	6.6
AX	118.6	113.4	IY	48.1	46.4	T	101.9	97.7
AY	50.6	48.4	JH	6.9	6.7	TH	10.0	10.1
B	24.0	22.8	K	47.3	44.4	UH	7.9	7.6
CH	4.9	4.7	L	49.1	47.1	UW	34.4	31.1
D	47.8	49.0	M	38.4	37.1	V	22.4	21.8
DH	42.9	40.0	N	90.0	85.5	W	27.6	27.1
DX	8.1	7.3	NG	16.4	16.5	Y	37.8	32.8
EH	33.6	32.5	OW	36.9	36.7	Z	30.3	29.1
ER	21.5	21.3	P	21.0	19.2			

Table B.2: *Frequency of occurrence in NIST SRE 2004, 2005 and 2006 conversations for every diphone-constraint.*

Average number of occurrences per conversation								
Diphone	Male	Female	Diphone	Male	Female	Diphone	Male	Female
AA-R	24.7	23.5	DH-AE	5.9	6.1	N-DH	4.0	4.1
AA-T	21.2	18.3	DH-AX	6.4	5.9	NG-K	4.2	4.4
AE-N	17.9	17.9	DH-EH	6.0	5.9	N-IY	3.7	3.5
AE-T	14.3	14.1	DH-EY	6.7	6.3	N-OW	4.2	4.0
AE-V	13.7	12.0	D-IH	6.0	5.0	N-S	4.4	3.5
AH-M	13.0	13.4	D-OW	5.8	6.2	N-T	3.6	3.9
AH-N	12.8	12.7	D-UW	5.5	5.8	OW-N	3.7	3.5
AH-T	11.9	9.9	DX-AX	5.8	5.7	P-AX	4.0	3.6
AO-L	11.2	11.3	DX-IY	6.1	5.4	R-AX	3.5	3.4

Continued on next page

Table B.2 – continued from previous page

Diphone	Male	Female	Diphone	Male	Female	Diphone	Male	Female
AO-R	9.8	9.0	EH-L	5.1	4.7	R-AY	3.6	4.0
AW-T	10.9	10.9	EH-N	5.5	5.1	R-IY	3.8	3.3
AX-B	11.9	11.0	EH-R	5.1	4.7	S-AH	3.7	3.7
AX-D	9.4	8.6	HH-AE	4.7	4.6	S-AX	3.5	3.2
AX-G	11.4	10.7	HH-W	4.9	4.8	S-OW	3.3	2.9
AX-K	10.2	9.0	IH-N	4.7	5.2	S-T	3.6	3.7
AX-L	9.6	9.4	IH-NG	5.4	4.8	T-AX	3.4	3.2
AX-M	8.9	8.7	IH-T	5.2	4.8	T-AY	3.7	3.5
AX-N	9.3	8.2	IY-AX	5.0	5.0	T-DH	3.1	3.6
AX-NG	8.8	8.2	IY-N	4.6	4.4	TH-IH	3.3	3.2
AX-S	8.2	7.7	IY-P	4.3	4.6	T-R	3.5	3.6
AX-T	8.0	7.9	JH-AX	4.3	4.2	T-S	3.4	3.5
AX-V	8.4	7.6	K-AH	4.3	4.3	T-UW	3.0	3.1
AX-Z	7.6	6.8	K-AX	4.4	4.2	T-W	3.2	2.9
AY-D	7.5	7.2	K-S	4.8	4.4	UH-D	3.0	3.0
AY-K	7.0	6.7	L-AX	4.4	3.9	UW-N	2.9	3.1
AY-M	7.5	7.1	L-AY	4.6	4.7	V-AX	3.2	3.0
AY-N	6.8	6.6	L-IY	4.2	4.4	W-AH	3.1	3.2
AY-T	7.4	6.4	M-AX	4.1	3.7	W-AX	3.2	2.9
B-AH	8.0	7.0	M-AY	4.5	3.7	W-EH	3.0	3.0
B-AX	6.1	6.1	M-IY	4.3	4.4	Y-AE	3.1	2.5
B-IY	6.3	6.5	N-AA	3.9	3.9	Y-UW	2.9	3.2
D-AX	6.6	6.2	N-AX	4.0	3.5	Z-AX	3.1	3.3
D-DH	6.4	6.5	N-D	4.2	3.8			

3.7. Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors

Título: "Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors"
Autores: Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez
Congreso: "Odyssey 2016: The Speaker and Language Recognition Workshop", 21-24 de Junio, 2016, Bilbao (España)
Pendiente de publicación

Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors

Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Universidad Autonoma de Madrid, Spain
javier.franco@uam.es

Abstract

In this paper, a probabilistic model is introduced to obtain feature-based likelihood ratios from linguistically-constrained formant-based i-vectors in a NIST SRE task. Linguistically-constrained formant-based i-vectors summarize both the static and dynamic information of formant frequencies in the occurrences of a given linguistic unit in a speech recording. In this work, a two-covariance model is applied to these "higher-level" features in order to obtain likelihood ratios through a probabilistic framework. While the performance of the individual linguistically-constrained systems are not comparable to that of a state-of-the-art cepstral-based system, calibration loss is low enough, providing informative likelihood ratios that can be directly used, for instance, in forensic applications. Furthermore, this procedure avoids the need for further calibration steps, which usually require additional datasets. Finally, the fusion of several linguistically-constrained systems greatly improves the overall performance, achieving very remarkable results for a system solely based on formant features. Testing on the English-only trials of the core condition of the NIST 2006 SRE (and using only NIST SRE 2004 and 2005 data for background and development, respectively), we report equal error rates of 8.47% and 9.88% for male and female speakers respectively, using only formant frequencies as speaker discriminative information.

1. Introduction

Formant frequencies have strong individualization potential [1] and have been used for forensic voice comparison for several decades [2]. However, most of the studies in automatic speaker recognition over the last two decades [3] have been based on higher dimensional representations of the speech signal (i.e. MFCC, PLP, etc.) due to their ability to extract speaker distinguishing information. Although they are based on spectral information, it is difficult to directly relate the physiological traits of an individual with the set of such extracted features [4]. Formant frequencies, on the other hand, are easily interpretable and directly related with anatomical and physiological characteristics [5] [1]. Moreover, interpretable features are helpful in order to correlate with human observations and may lead to find some clues that could be hidden even for very complex cepstral-based systems [6].

In forensic-phonetics, voice comparison is usually performed in the context of linguistic units [1, 7], but reported studies are usually based on limited experimental frameworks (in terms of number of speakers, number of analysed linguistic-units, or both) due to the manual processes involved in order to extract formant frequencies or labelling the analysed units. So, it is of broad interest to analyse the abilities of formant frequencies for speaker recognition following a similar approach but applied on a large-scale experimental framework with the aid of fully automatic systems.

While there have been previous studies on the use of formant frequencies for automatic speaker recognition [8], constraints have been used only in the feature extraction stage but not for speaker modelling. On the other hand, there have been several studies on text-constrained speaker modelling but using mainly cepstral [9] or prosodic features [10]. In both cases, forensic applications have not been addressed in depth. Thus, to some extent this research fill a gap in the literature, and the presented results can give useful insights for the practitioners in the forensic-phonetics field.

In [11], the authors showed that well calibrated likelihood ratios can be obtained per linguistic unit by means of i-vector systems independently developed from linguistically-constrained formant features. However, as using a simple scoring method, an additional calibration step was needed in order to obtain informative likelihood ratios. In this work, a probabilistic framework is applied instead, leading to likelihood ratios that can be directly used avoiding further calibration processes, which usually need additional datasets in order to avoid overoptimistic results. This probabilistic framework is based on a two-covariance generative model similar to that in [12], but with a simpler training step as it has been used in some forensic works [13, 14].

The remainder of the paper is organized as follows. The extraction process of linguistically-constrained formant-based i-vectors is detailed in Section 2. Section 3 introduces the probabilistic model applied to the linguistically-constrained formant-based i-vectors in order to obtain feature-based likelihood ratios. Section 4 describes the experimental framework used for this work, while Section 5 presents the results obtained. Finally, conclusions are drawn in Section 6.

2. Linguistically-constrained formant-based i-vectors

Linguistically-constrained formant-based i-vectors are extracted with the aid of several speech processing tools and attempt to summarize both the static and dynamic information of formant frequencies in the occurrences of a given linguistic unit in a speech recording. First, automatic formant tracking is used in order to obtain the formant frequencies in a given speech file. In order to account for the dynamic information, delta features are also computed and incorporated to the feature vectors. Then, an automatic speech recognition (ASR) system is used to split the stream of feature vectors into different linguistic units.

Finally, for each speech recording, the feature vectors corresponding to the occurrences of a given linguistic unit are used to compute a linguistically-constrained i-vector for that utterance.

2.1. Formant tracking and dynamic information

Automatic formant tracking has been used in order to compute the formant frequencies along a speech recording. Among the free software packages available, Wavesurfer [15] has been selected for this work due to the ease of automate this process for large databases through scripts written in Tcl/Tk [16], as it is developed using the Snack Sound Toolkit library [17].

The Wavesurfer/Snack formant tracker bases its formant-frequency estimates on a linear prediction analysis performed at each frame, and dynamic programming is used to refine the resulting trajectories [18]. It has been used with default parameters for both male and female speakers, except for the number of formant frequencies to be tracked, limited to three for this work (F1-F3).

In order to account for the dynamic information of formant frequencies, the *delta* (Δ) or derivative coefficients have been used. Although delta coefficients cannot summarize the whole formant trajectory along a linguistic segment as other approaches attempt [19, 20, 21], they can characterize the local dynamic information while keeping a frame-by-frame rate and a low dimensionality [11]. Derivative coefficients are finally appended to the instantaneous formant frequencies at each frame (10 ms each), giving rise to our 6-dimensional lower-level feature vectors.

2.2. Region conditioning and types of constraints

Voice comparison in forensic-phonetics is usually performed in the context of linguistic units [1, 7], as formant frequencies present much lower intra-speaker variability and higher inter-speaker variability [22, 7] when these constraints are applied to the features to be compared.

Automatic speech recognition (ASR) systems provide both phonetic content and time interval of speech regions in which the audio stream can be segmented. This phonetic content allow to define a large set of candidate constraints among the different types of linguistic units, showing each of them different characteristics in terms of within-unit formant dynamics, unit-length and frequency of occurrence. Among them, the following have been used for this work:

- Phones: although they are the shortest units and can appear in many different linguistic contexts, their high frequency of occurrence allow to develop more robust constrained systems. For this work, 39 phone units from an English lexicon plus two filled pauses (represented as PUH and PUM) were selected. These linguistic units are represented by the "2-character" ARPABET symbols [23] in the phonetic transcriptions provided by the ASR system [24] used. Table 1 shows the correspondence between Arpabet symbols and the International Phonetic Alphabet (IPA) ones.
- Diphones: defined as every possible combination of phone pairs, the 98 most frequent diphones were selected. Compared with phones, they present longer length but much lower frequency of occurrence. However, they show less contextual variation, which may lead to reduce the intra-speaker variability of formant dynamics between different occurrences of the same diphone.

In this work, the phonetic transcription labels produced by the SRIs Decipher ASR system [24] are used. For this system, trained on English data from telephonic conversations, the Word Error Rate (WER) on native and non-native speakers on the transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively.

2.3. I-vector extraction

An i-vector extractor [25] is a factor analysis (FA) based front-end which attempts to summarize the speaker distinguishing information in a given utterance, represented by a set of L feature vectors $\{f_1, f_2, ..., f_L\}$, through a single low-dimensional vector, the so-called identity vector or *i-vector* for short. This i-vector w accounts for the speaker and channel/session information present in a given utterance, representing it in a low-dimensional variability subspace. This is done by converting the speaker- and session-independent supervector (m), usually taken to be the UBM supervector, into the speaker- and session-dependent supervector (M) that represents a given speaker utterance through:

$$M = m + Tw \tag{1}$$

where T is a rectangular matrix of low rank defining the total variability (TV) space that contains the speaker and channel variability.

In order to obtain a linguistically-constrained i-vector (w^C), the i-vector extractor is applied only to the set of feature vectors $\{f_1^C, f_2^C, ...\}$ in the utterance belonging to a particular constraint, C:

$$M^C = m^C + T^C w^C \tag{2}$$

For this purpose, independent UBMs and TV subspaces are trained on the background dataset (see Section 4 for details) for every linguistic constraint under analysis. Both the number of components of the UBM (ranging from 2 to 256) and the number of dimensions of the TV space (ranging from 5 to 50) are optimized on the development dataset (see Section 4 for details) for each linguistic unit/constraint.

Finally, linguistically-constrained i-vectors are centred and whitened on the background dataset, and length-normalized.

3. Probabilistic model

3.1. The generative model

Conversely to [11], were cosine scoring and a further calibration step were used, in this work likelihood ratios are directly derived through a probabilistic framework. For this purpose, a two-covariance model [12] is applied. This is a generative model in which a particular observed i-vector $\mathbf{x}_{ij}$ coming from speaker i is generated through

$$\mathbf{x}_{ij} = \boldsymbol{\theta}_i + \boldsymbol{\psi}_j \tag{3}$$

where $\boldsymbol{\theta}_i$ is a realization of the speaker random variable Θ and $\boldsymbol{\psi}_j$ is a realization of the additive random noise Ψ representing its within-speaker variation. This noisy term is taken to be constant among different speakers and randomly distributed following

$$\Psi \sim \mathcal{N}(0, \mathbf{W}) \tag{4}$$

where $\mathbf{W}$ is the within-speaker covariance matrix. Thus, the conditional distribution of the random variable X_i (from which

$\mathbf{x}_{ij}$ is drawn), given a particular speaker i, follows a normal distribution with mean $\boldsymbol{\theta}_i$ and covariance matrix $\mathbf{W}$

$$\mathrm{X}_{ij}|(\Theta = \boldsymbol{\theta}_i) \sim \mathcal{N}(\boldsymbol{\theta}_i, \mathbf{W}) \tag{5}$$

On the other hand, speakers means are assumed to be normally distributed, following

$$\Theta \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}) \tag{6}$$

where $\boldsymbol{\mu}$ and $\mathbf{B}$ are, respectively, the mean vector and the covariance matrix of the between-speaker distribution.

3.2. Model training

Conversely to [12], model *hyperparameters* are directly computed in a single step instead of being iteratively trained to maximize the likelihood of the true partitioning of *m* speakers in the background dataset. This alternative procedure is more commonly used in forensic studies [13, 14], and it is applied in this work in order to avoid overfitting to the limited background dataset (NIST 2004 SRE).

Within-speaker covariance matrix is computed from the background dataset $\mathbf{X}$, comprising N i-vectors coming from m different speakers, through

$$\mathbf{W} = \frac{\mathbf{S}_w}{N - m} \tag{7}$$

being $\mathbf{S}_w$ the within-speaker scatter matrix given by

$$\mathbf{S}_w = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathrm{T}} \tag{8}$$

where $\bar{\mathbf{x}}_i$ is the average of the set of n_i i-vectors from speaker i.

On the other hand, the mean vector and the covariance matrix of the between-speaker distribution are respectively computed by

$$\boldsymbol{\mu} = \frac{1}{m} \sum_{i=1}^{m} \bar{\mathbf{x}}_i \tag{9}$$

and

$$\mathbf{B} = \frac{\mathbf{S}_b}{m - 1} - \frac{\mathbf{S}_w}{\bar{n}(N - m)} \tag{10}$$

where $\bar{n}$ is the average number of i-vectors per speaker and the between-speaker scatter matrix, $\mathbf{S}_b$, is given by

$$\mathbf{S}_b = \sum_{i=1}^{m} (\bar{\mathbf{x}}_i - \boldsymbol{\mu})(\bar{\mathbf{x}}_i - \boldsymbol{\mu})^{\mathrm{T}} \tag{11}$$

3.3. Likelihood-ratio computation

Finally, the likelihood ratio between two given linguistically-constrained i-vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ is computed as the ratio between

$$p(\mathbf{y}_1, \mathbf{y}_2) = \int_{\boldsymbol{\theta}} p(\mathbf{y}_1|\boldsymbol{\theta}, \mathbf{W})\, p(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{W})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \tag{12}$$

and

$$p(\mathbf{y}_1) \cdot p(\mathbf{y}_2) = \int_{\boldsymbol{\theta}} p(\mathbf{y}_1|\boldsymbol{\theta}, \mathbf{W})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \times \int_{\boldsymbol{\theta}} p(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{W})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \tag{13}$$

where $p(\mathbf{y}_l|\boldsymbol{\theta}, \mathbf{W}) = N(\mathbf{y}_l; \boldsymbol{\theta}, \mathbf{W})$ is the probability of a linguistically-constrained i-vector $\mathbf{y}_l$ given the knowledge of the speaker $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\mathbf{X}) = N(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{B})$ is the between-speaker probability density function obtained from the background dataset $\mathbf{X}$. Closed form expressions for these integrals can be found, for example, in [26].

4. Experimental framework

4.1. Datasets

In order to develop and test the linguistically-constrained systems, we have used the datasets and protocols belonging to the NIST SREs carried out on years 2004 [27], 2005 [28] and 2006 [29]. Among them, only English conversations have been used in order to match the characteristics of the ASR system [24]. No other datasets have been used as the authors have access only to the ASR phonetic labels corresponding to those datasets, kindly provided by SRI.

The composition of these datasets and the purposes they have been devoted to are described below:

- Background: NIST 2004 SRE dataset [27] comprises 2,541 files (1378 5-minutes, 581 30-seconds and 582 10-seconds long) from 125 male speakers and 3,626 files (2022 5-minutes, 802 30-seconds and 802 10-seconds long) from 187 female speakers. It has been used as the background dataset for training UBMs and total variability matrices. It also has been reused in order to train the *hyperparameters* of the probabilistic model.
- Development: NIST 2005 SRE dataset [28] has been used in order to optimize both UBMs components and number of dimensions of the TV subspaces. In [11], this dataset was divided into two halves for additional purposes: one half was used to train the calibration process and the other one to train the fusion rules. Here, as the calibration step is avoided through the introduced probabilistic framework, the whole dataset is used to train the fusion rules. The 1side-1side task of this NIST SRE comprises 11,272 trials from 243 male speakers and 14,793 trials from 342 female speakers.
- Evaluation: English-only trials from the core condition of the NIST 2006 SRE [29] were used for evaluating the proposed approach, consisting of 9,720 male trials for 219 target speakers and 14,293 female trials for 298 target speakers.

4.2. Evaluation metrics

Both the calibration and the discriminative properties of linguistically-constrained systems are analysed in this work. Discriminative properties are mainly evaluated through the equal error rate (EER) [30]. It is also used as the criterion by which the subsets of constraints are selected for the combination of linguistically-constrained systems. However, in accordance to the protocols used [29], the minimum of the C_{Det} (minDCF) is also reported. On the other hand, calibration properties [31] of linguistically-constrained systems are evaluated through the C_{llr} cost function and the calibration loss (C_{llr}^{loss}) [32].

5. Results

5.1. Reference systems

First, we want to compare with the previous approach in [11], were the same linguistically-constrained formant-based i-

vectors were used. In that study, cosine scoring, z-norm and a calibration step were used to obtain LRs per linguistic-unit. The results per constraint for this system can be seen in Table 2 (only the 10 best performing constraints are shown). The best fused system was obtained through a logistic regression fusion of the N-best performing constraints (see Section 5.3 or [11] for more details), trained on the same development dataset used in this work (NIST 2005 SRE). The performance of this fused system on the evaluation dataset (NIST 2006 SRE) is shown in Table 1.

Secondly, we wan to compare with a state-of-the-art cepstral-based system. Our cepstral reference system is based on an i-vector extractor from (unconstrained) MFCC features [25] and a Gaussian PLDA scoring stage [33]. Both gender-dependent 1024-component UBMs and 600-dimensional TV subspaces are trained on the background dataset (NIST SRE 2004), but GPLDA *hyperparameters* are trained on both the background and the development dataset (NIST SRE 2004 and 2005), applying a dimensionality reduction to 200. The performance of this system on the evaluation dataset (NIST SRE 2006) is also shown in Table 1.

	Reference systems			
	Male		Female	
	EER (%)	minDCF	EER (%)	minDCF
Formant-based	9.57	0.0456	12.89	0.0543
Cepstral-based	4.21	0.0232	5.67	0.0303

Table 1: *Results on the evaluation dataset for both formant-based and cepstral-based reference systems.*

5.2. Independent linguistically-constrained systems

Table 3 shows the results for the 10-best performing constraints (in terms of the EER) on the evaluation dataset when the introduced probabilistic framework is applied. As it can be seen, compared to the previous approach (Table 2), the discriminative performance per linguistic-unit (in terms of the EER) is significantly improved ($\sim$15% relative improvement on average for the shared constraints among the 10-best performing ones). Furthermore, although it is slightly increased compared to the previous approach, very low calibration losses are obtained (C_{llr}^{loss} $\sim$0.04 on average for this 10-best performing set) without the need for a specific calibration step.

5.3. Fusion of linguistically-constrained systems

Feature-based likelihood-ratios from different linguistic-constraints can be combined in order to account for the speaker distinguishing information spread among the different units. In this work, two fusion techniques have been used:

- First, a simple fusion rule, consisting on averaging the log-LRs of the subset of *N* constraints to be combined, has been applied through

$$logLR = \frac{1}{N} \sum_{\forall C \; in \; subset} logLR^C \qquad (14)$$

  where $logLR^C$ is the log-LR for a particular constraint *C*.

- Secondly, a linear combination of log-LRs is applied through

$$logLR = \alpha_0 + \sum_{\forall C \; in \; subset} \alpha^C logLR^C \qquad (15)$$

  where the vector of weights $\boldsymbol{\alpha} = [\alpha_0, \alpha^{C_1}, \alpha^{C_2}, ..., \alpha^{C_N}]$ is obtained by *logistic regression* [34] training on the development database, using the FoCal toolkit [35].

The specific subset of *N* constraints to be fused is obtained as follows. First, linguistically-constrained systems are sorted by performance, in terms of the EER, on the development dataset. Then, different fused systems are obtained by combining the first two, three, *etc.*, best performing linguistically-constrained systems, up to the total number of constraints. Among them, the fused system with the best performance on the development dataset, which is obtained by fusing the *N*-best performing constraints, is selected. While this may not be the set of constraints with the best performance on the evaluation dataset, it is expected that, for a large enough value of N, most of the best-performing constraints will be shared among development and evaluation datasets.

Table 4 shows the results obtained for both fusion techniques on the evaluation dataset. As it can be seen, the average rule make use of a much lower number of constraints than the logistic regression technique. This issue was analysed in [11], where it was shown that the performance of the logistic regression technique on the development dataset improved as more constraints were fused. For male trials, while the discriminative capabilities are similar for both techniques in terms of EER and minDCF, calibration properties are significantly better for the logistic regression technique, specially the calibration loss. The latter is also true for female trials, but also the discriminative capabilities are significantly better compared to the average rule.

Regarding our reference systems (Table 1), the logistic regression fusion of the feature-based LRs obtains a relative improvement in performance, in terms of the EER, of 11.5% and 23.3% for male and female trials, respectively, compared to our formant-based reference. While the performance is still far from the cepstral-based reference system, it is a very remarkable result for a system solely based on formant features which, in addition, can be directly applied in different forensic settings.

6. Conclusions

In this paper, we have introduced a probabilistic framework in order to obtain feature-based likelihood ratios from formant-based linguistically-constrained i-vectors. Linguistically-constrained formant-based i-vectors summarize both the static and dynamic information of formant frequencies in the occurrences of a given linguistic unit in a speech recording, and are extracted in a fully automatic way with the aid of several speech processing tools, including automatic formant tracking and automatic speech recognition.

A probabilistic model applied to formant-based i-vectors of a given linguistic constraint allows to provide feature-based likelihood ratios for isolated linguistic units, avoiding further calibration processes which usually need additional datasets. Although the discriminative performance of linguistically-constrained systems is not comparable to that of a cepstral-based state-of-the-art system, informative calibrated LRs can be obtained for voice comparisons without the need for further calibration steps.

Cosine scoring + z-normalization + calibration (log. reg.)									
Male					Female				
Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AE	21.21	0.0850	0.6668	0.0143	AY	24.59	0.0841	0.7101	0.0111
AY	21.38	0.0825	0.6580	0.0158	AE	24.59	0.0876	0.7308	0.0131
N	22.26	0.0812	0.6896	0.0168	L	24.68	0.0869	0.7355	0.0127
L	23.24	0.0839	0.7083	0.0133	N	24.77	0.0839	0.7256	0.0112
AX	23.80	0.0844	0.7001	0.0150	R	26.49	0.0932	0.7681	0.0132
AH	23.96	0.0964	0.7286	0.0158	AX	27.15	0.0932	0.7764	0.0100
PUH	24.32	0.0933	0.7296	0.0137	OW	27.79	0.0936	0.7830	0.0098
Y	24.68	0.0915	0.7325	0.0180	DH	27.79	0.0940	0.7876	0.0114
EH	24.83	0.0972	0.7544	0.0140	EH	28.06	0.0990	0.8196	0.0157
R	24.96	0.0937	0.7380	0.0149	AH	28.89	0.0974	0.8185	0.0079

Table 2: *Results on the evaluation dataset for the 10 best-performing constraints obtained with the previous approach in [11].*

Two covariance model									
Male					Female				
Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
L	18.21	0.0817	0.6074	0.0421	L	20.21	0.0833	0.6542	0.0360
N	18.35	0.0764	0.6373	0.0574	N	21.02	0.0805	0.7455	0.1190
AY	18.43	0.0813	0.6090	0.0318	AE	21.16	0.0849	0.6896	0.0521
AE	19.50	0.0835	0.6411	0.0390	AY	21.58	0.0814	0.6606	0.0261
R	20.61	0.0889	0.6672	0.0275	AX	23.16	0.0898	0.6937	0.0144
Y	21.38	0.0888	0.6918	0.0575	R	23.59	0.0913	0.7147	0.0240
AX	21.50	0.0830	0.7012	0.0460	DH	24.30	0.0925	0.7316	0.0320
IH	21.76	0.0926	0.6885	0.0298	AH	25.05	0.0952	0.7541	0.0209
OW	21.90	0.0899	0.6911	0.0284	Y-AE	25.15	0.0866	0.7976	0.0914
DH	22.32	0.0892	0.6739	0.0231	OW	25.33	0.0906	0.7265	0.0294

Table 3: *Results on the evaluation dataset for the 10 best-performing constraints when the introduced probabilistic framework is applied.*

	***N*-best fusion of linguistically-constrained systems**									
	Male					Female				
	N	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	*N*	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
Average rule	19	8.76	0.0444	0.4318	0.1342	15	11.82	0.0565	0.4743	0.0864
Logistic regression	138	**8.47**	**0.0451**	**0.3210**	**0.0269**	139	**9.88**	**0.0512**	**0.3488**	**0.0140**

Table 4: *Results on the evaluation dataset for the fusion of the N-best performing linguistically-constrained systems, for the average rule and the logistic regression fusions.*

Furthermore, feature-based LRs can be successfully combined through different fusion techniques, obtaining great improvements in discriminative performance compared with the independent linguistically-constrained systems by themselves. For a simple average fusion rule, tens of units can be fused at the cost of slightly higher calibration losses. For the logistic regression technique, as being a trained fusion rule, a larger number of units can be fused while keeping very good calibration properties. While the performance is still far from the cepstral-based reference system, it is a very remarkable result for a system solely based on formant features which, in addition, can be directly applied in different forensic settings.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz, TEC2012-37585-C02-01). Also, the authors would like to thank SRI for providing the Decipher phonetic transcriptions of the NIST 2004, 2005 and 2006 SREs that have allowed to carry out this work.

8. References

[1] Phillip Rose, *Forensic Speaker Identification*, Forensic Science. Taylor and Francis, 2002.

[2] Francis Nolan, *The phonetic bases of speaker recognition*, Cambridge University Press, Cambridge (UK), 1983.

[3] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] Jonathan Darch, Ben Milner, Xu Shao, Saeed Vaseghi, and Qin Yan, "Predicting formant frequencies from MFCC vectors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 2005, pp. 941–944.

[5] Francis Nolan and Catalin Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law*, vol. 12, no. 2, 2005.

[6] Joaquin Gonzalez-Rodriguez, Juana Gil, Rubén Pérez, and Javier Franco-Pedroso, "What are we missing with i-vectors? a perceptual analysis of i-vector-based falsely accepted trials," in *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 33–40.

[7] Kirsty McDougall, "Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies," *International Journal of Speech Language and the Law*, vol. 13, no. 1, pp. 89 – 126, 2006.

[8] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, Sept 2007.

[9] Tobias Bocklet and Elizabeth Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4525–4528.

[10] Elizabeth Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I. Fundamentals, Features, and Methods*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 241–259, Springer Berlin Heidelberg.

[11] Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61 – 81, 2016.

[12] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 34.

[13] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," vol. 53, no. 1, pp. 109–122, Feb. 2004.

[14] Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, and Colin Aitken, *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data.*, Wiley, 2014.

[15] Kåre Sjölander and Jonas Beskow, "Wavesurfer - an open source speech tool," in *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000 / INTERSPEECH), Beijing, China, October 16-20, 2000*, 2000, pp. 464–467.

[16] "Tcl — Wikipedia, The Free Encyclopedia," 2015.

[17] "Snack Sound Toolkit — Wikipedia, The Free Encyclopedia," 2014.

[18] "Snack v2.2.8 manual," .

[19] Najim Dehak, Patrick Kenny, and Pierre Dumouchel, "Continuous prosodic features and formant modeling with joint factor analysis for speaker verification," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 1234–1237.

[20] Joaquin Gonzalez-Rodriguez, "Speaker recognition using temporal contours in linguistic units: The case of formant and formant-bandwidth trajectories," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy, August 27-31, 2011*, 2011, pp. 133–136.

[21] Javier Franco-Pedroso, Fernando Espinoza-Cuadros, and Joaquin Gonzalez-Rodriguez, "Formant trajectories in linguistic units for text-independent speaker recognition," in *Proceedings of the International Conference on Biometrics (ICB 2013), 4-7 June, 2013, Madrid, Spain*, 2013, pp. 1–6.

[22] Francis Nolan, "The 'telephone effect' on formants: a response," *International Journal of Speech Language and the Law*, vol. 9, no. 1, 2002.

[23] J. E. Shoup, "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, Wayne A. Lea, Ed. 1980, pp. 125–138, Englewood Cliffs: Prentice Hall.

[24] Sachin S. Kajarekar, Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer, and Tobias Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4205–4208.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[26] Javier Franco-Pedroso, Daniel Ramos, and Joaquin Gonzalez-Rodriguez, "Gaussian mixture models of between-source variation for likelihood ratio computation from multivariate data," *PLoS ONE*, vol. 11, no. 2, pp. 1–25, 02 2016.

[27] "The NIST Year 2004 Speaker Recognition Evaluation Plan," .

[28] "The NIST Year 2005 Speaker Recognition Evaluation Plan," .

[29] "The NIST Year 2006 Speaker Recognition Evaluation Plan," .

[30] Joaquin Gonzalez-Rodriguez, "Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014)," *Loquens*, vol. 1, no. 1, pp. 1–15, January 2014.

[31] DavidA. van Leeuwen and Niko Brümmer, "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems," in *Speaker Classification I*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in Computer Science*, pp. 330–353. Springer Berlin Heidelberg, 2007.

[32] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," in *Computer Speech and Language*, 2006, vol. 20, pp. 230 – 275.

[33] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252.

[34] Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.

[35] Niko Brümmer, "Toolkit for evaluation, fusion and calibration of statistical pattern recognizers," .

[36] "Arpabet — Wikipedia, The Free Encyclopedia," 2014.

A. Mathematical notation

Column vectors are denoted by bold lower-case letters and matrices by bold upper-case letters, while scalar quantities are denoted by lower-case italic letters. Random variables are denoted by upper-case non-italic letters. $P(\cdot)$ is used to indicate the probability of a certain event, while $p(\cdot)$ denotes a probability density function. We denote a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the corresponding probability density function by $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\mathbf{x} \in \mathbb{R}^d$).

B. Phonetic transcription codes

Vowels					
Monophthongs			*Monophthongs*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
AO	ɔ	off; fall; frost	AE	æ	at; fast
AA	ɑ	father; cot	*Diphthongs*		
IY	i	bee; see	Arpabet	IPA	Word examples
UW	u	you; new; food	EY	eɪ	say; eight
EH	ɛ	red; men	AY	aɪ	my; why; ride
IH	ɪ	big; win	OW	oʊ	show; coat
UH	ʊ	should; could	AW	aʊ	how; now
AH	ʌ	but; sun	*R-coloured vowels*		
	ə	sofa; alone	Arpabet	IPA	Word examples
AX		discus	ER	ɝ	her; bird; heart; nurse

Consonants					
Stops			*Affricates*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
P	p	pay	CH	tʃ	chair
B	b	buy	JH	dʒ	just
T	t	take	*Semivowels*		
D	d	day	Arpabet	IPA	Word examples
K	k	key	Y	j	yes
G	g	go	W	w	way
Fricatives			*Liquids*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
F	f	for	L	ɬ	late
V	v	very	R	ɹ *or* ɻ	run
TH	θ	thanks; Thursday	DX	ɾ	wetter
DH	ð	that; the; them	*Nasals*		
S	s	say	Arpabet	IPA	Word examples
Z	z	zoo	M	m	man
SH	ʃ	show	N	n	no
HH	h	house	NG	ŋ	sing

Table 1: *39 phones from the Arpabet phonetic transcription code and their correspondent IPA symbols (extracted from [36]).*