

REPORTE TÉCNICO
**Reconocimiento
de Patrones**

**Métodos de reconocimiento
automático del idioma**

Ing. Oneisys Núñez Cuadrta,
Dr. C. José Ramón Calvo de Lara

RT_015

Noviembre 2009





CENATAV

Centro de Aplicaciones de
Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
**Reconocimiento
de Patrones**

**Métodos de reconocimiento
automático del idioma**

Ing. Oneisys Núñez Cuadra,
Dr. C. José Ramón Calvo de Lara

RT_015

Noviembre 2009



Métodos de reconocimiento automático del idioma

Ing. Oneisys Núñez Cuadra, Dr. C. José Ramón Calvo de Lara

Centro de Aplicaciones de Tecnología de Avanzada, 7a #21812 e/ 218 y 222, Siboney, Playa, Habana, Cuba
onunez@cenatav.co.cu

RT_015 CENATAV

Fecha del camera ready: 28 de agosto de 2009

Resumen: El reconocimiento o identificación del idioma en que una persona está hablando es una de las primeras tareas a las que tiene que enfrentarse cualquier sistema automático que interactúe directamente con usuarios multilingües. Hoy en día, se ha desarrollado toda una amplia gama de algoritmos y métodos que permiten identificar en qué idioma está una expresión dada. Con el objetivo de desarrollar las tecnologías de identificación del lenguaje (LID) los investigadores han abierto varios enfoques, cada uno en función de tener un mejor aprovechamiento de los datos y mejores resultados en la tarea en cuestión. El trabajo propuesto abarca un estado del arte de los diferentes métodos de identificación del idioma y fundamentalmente de los sistemas LID que se han desarrollado hasta el momento.

Palabras claves: reconocimiento del idioma, sistemas LID

Abstract: The language recognition or identification, in which a person is talking, is one of the first tasks that must face any automatic system that interacts directly with multilingual users. Today, it has developed a wide range of algorithms and methods to identify in which language is a given expression. In order to develop technologies for language identification (LID) researchers have opened several approaches, each based on a better use of data and better results on the task at hand. The proposed work encompasses a state of art of different methods of language identification and essentially of the LID systems have been developed so far.

Keywords: Language Recognition, LID systems

1 Introducción

La identificación del lenguaje hablado consiste en determinar el idioma de quien habla basándose sólo en una muestra de voz sin considerar al hablante y lo que está diciendo. De acuerdo a esto, la identificación automática del lenguaje hablado (LID) es el proceso por el cual el lenguaje (idioma) de una muestra de señal de voz digitalizada es reconocido por una computadora [1].

En la actualidad existen un conjunto grande de sistemas cuyo resultados finales dependen en gran medida de esta tecnología LID, constituyendo un modulo importante dentro de varias aplicaciones como sistemas de conversación multilingüe [2] traducción de lenguaje hablado [3], reconocimiento del habla multilingüe [4], recuperación de documentos hablados [5]. También es un asunto de gran importancia en las áreas de inteligencia y seguridad donde se necesita conocer la identificación del idioma de mensajes grabados y documentos archivados antes de poder extraer cualquier información de ellos. Otra aplicación en esta rama seria la determinación del origen del hablante a partir del reconocimiento de su lenguaje autóctono.

Desde la perspectiva del reconocimiento del idioma, los seres humanos están constantemente expuestos a ellos y sin que les sea dado ningún conocimiento aprenden a determinar la identidad

de los lenguajes detectando algunas especificidades del habla en los diferentes idiomas. Por ejemplo una persona que habla inglés puede apreciar el tono natural del idioma Chino-Mandarín. En los experimentos de percepción humana, cuando los oyentes son sometidos a la identificación de idiomas poco familiares a ellos, aquellos con una base multilingüe obtienen mejores resultados que los que conocen un solo idioma [6]. El conocimiento léxico es generalmente la principal fuente de información que tienen los humanos, especialmente cuando los idiomas a distinguir son similares. Estas fueron, entre otras, las razones que motivaron a explorar la utilidad de los atributos del habla en el idioma, siguiendo la misma línea del paradigma LID propuesta recientemente para el reconocimiento automático del habla (ASR) [7]. Experimentos como estos, mostraron que el contenido prosódico y silábico son características del habla, y del lenguaje, muy útiles y que por tanto, podían ser incorporadas dentro de los vectores de rasgos.

Por su parte las máquinas están equipadas para tomar virtualmente una cantidad ilimitada de expresiones y textos, por lo tanto, su potencial sobrepasa el funcionamiento humano para el reconocimiento del lenguaje hablado. Los sistemas preparados para asumir esta tarea, a menudo consideran las expresiones desconocidas como una concatenación de patrones de la señal, y se valen de modelos probabilísticos y técnicas de clasificación para caracterizar los patrones y sus correspondientes identidades. En estos casos los modelos léxicos no son explícitamente utilizados, solo en algunos sistemas donde se modelan los fonemas de un subconjunto de lenguajes y se usan para decodificar la señal en vista de poder realizar un mejor procesamiento.

En la actualidad existe una amplia colección de datos, conformados por ejemplos de expresiones de un conjunto de lenguajes seleccionados, gracias a los esfuerzos coordinados entre las agencias financiadas por los gobiernos y las comunidades técnicas. La disponibilidad de estos, el acceso a equipos de cómputo rápidos y fácilmente costeables, y los avances recientes en el modelado del lenguaje y en las máquinas de aprendizaje han propiciado la exploración de nuevos algoritmos en esta área; lo que ha provocado un rápido progreso en la rama de la modelación del lenguaje hablado y en las tecnologías de reconocimiento.

Un primer punto en la caracterización del lenguaje hablado es definir las fuentes de conocimiento que se van utilizar para distinguir un idioma de otro. Los contrastes entre el proceso de reconocimiento del lenguaje por los humanos y por las máquinas tienen como punto de interés que las pistas percibidas por los oyentes humanos puedan ser incorporadas a las tecnologías actuales para mejorar las potencialidades de las máquinas.

No se puede obviar el hecho de que existen muchos más lenguajes y dialectos que sistemas equipados para manipularlos. Muchos de esos lenguajes son raros y no se rigen por reglas lingüísticas, por lo que construir reconocedores para esos lenguajes resulta prohibitivo. Algunas de las tecnologías predominantes de hoy deben ser modificadas para aliviar estas limitaciones. Por ello se han establecido nuevos paradigmas que permitan examinar el problema de reconocimiento del lenguaje desde una perspectiva completamente diferente.

En la actualidad se han desarrollado diferentes tendencias para enfrentar este problema. Una de ellas, se enfoca en los métodos basados en espectros, en el que cada expresión hablada o segmento es representado por una secuencia de vectores de características que es a menudo una pequeña representación espectral de un fragmento de la expresión. Esto es semejante a lo que se hace generalmente en el módulo de extracción de características en los sistemas automáticos de reconocimiento de locutores. Se asume que estos vectores espectrales fueron generados desde fuentes de diferentes idiomas y por lo tanto tienen diferentes características. Los modelos, considerando uno para cada lenguaje, pueden ser contruidos etiquetando los ejemplos de

vectores espectrales tomados de todos los lenguajes. Ese modelo puede ser probabilístico como el Modelo de Mezclas Gaussianas GMM [8], o basado en vectores, como las Máquinas de Vectores de Soporte, SVM [9].

Como se puede ver todo este trabajo basado en espectros es puramente acústico, mientras que la información lingüística, como los fonemas o palabras, no son usadas. El segundo método es el llamado aproximación basado en tokens, en el que intervienen un conjunto de unidades del habla también referidas como tokens usadas para representar el discurso. Una expresión es primero decodificada y segmentada en secuencias semejantes a tokens. Este flujo de tokens es usado para extraer características, diseñar clasificadores y actuar como un proceso de clasificación de lenguaje hablado.

Otra tendencia desarrollada muy recientemente en el reconocimiento del lenguaje es la caracterización por vectores de espacio (VSC), en la cual una expresión es considerada como un documento hablado. El término documento es una representación matricial de todas las expresiones en el conjunto de entrenamiento que pueden ser establecidas para propósitos de indexación y de recuperación. Los términos aquí se refieren a las palabras acústicas (AWs), que son secuencias de letras acústicas (alfabeto) o unidades fundamentales de sonido, y los elementos del vector son a menudo las co-ocurrencias estadísticas que describen la frecuencia de las AW en cada expresión hablada. El reconocimiento del idioma para una entrada de habla desconocida puede ser tratado como un problema de recuperación de documentos similar a los que son comúnmente confrontados en la comunidad de recuperación de información. Esto también puede proyectarse como un problema de categorización de texto en el que todos los documentos de entrenamiento son usados para construir los categorizadores de texto o clasificadores de asuntos, uno por cada idioma considerado. De esta forma se puede finalizar el reconocimiento del lenguaje comparando el vector entrante de habla desconocida con cada uno de los clasificadores de lenguajes para construir una decisión de reconocimiento. Este modelado basado en vectores permite representar las expresiones habladas como un vector documento con una gran dimensionalidad, a veces en el orden de las miles, entonces la mayoría de los algoritmos de extracción de características y clasificación actualmente disponibles en la literatura para recuperación de información y categorización de texto pueden ser fácilmente adaptados para el reconocimiento del lenguaje hablado. Las pistas percibidas de los lenguajes también pueden ser exploradas e incorporadas dentro del vector del documento hablado, siguiendo la misma línea del paradigma para la transcripción automática de atributos del habla propuesta recientemente para el Reconocimiento automático del habla con el fin de mejorar el funcionamiento de los sistemas.

2 Identificación automática del lenguaje hablado

Las investigaciones para la identificación automática del lenguaje hablado se han desarrollado siguiendo diferentes enfoques. Sin embargo, todos ellos tienen en forma general dos fases para asumir esta laboriosa tarea, las que se ilustran en la figura 1.

Durante la fase de entrenamiento, el típico sistema es presentado con ejemplos de muestras de voz de una gran variedad de lenguajes. Cada muestra de entrenamiento es convertida a una cadena de vectores de características. Cada vector de características proviene de ventanas de la

señal digitalizada (por ejemplo ventanas cortas de 20ms) durante los cuales la señal de voz se asume cuasiestacionaria. Los vectores de características contienen información espectral o cepstral acerca de la señal de voz. El algoritmo de entrenamiento analiza una secuencia de dichos vectores y produce uno o más modelos para cada lenguaje. Esos modelos capturan las características propias del lenguaje. Durante la fase de reconocimiento, el vector de características procesado de una nueva señal de voz es comparada con cada modelo. Después se calcula la probabilidad de que esta nueva señal *pertenezca* a alguno de los modelos obtenidos al entrenar. Este cálculo se repite para cada uno de los lenguajes entrenados y finalmente se considera el lenguaje de la muestra aquel con la máxima probabilidad.

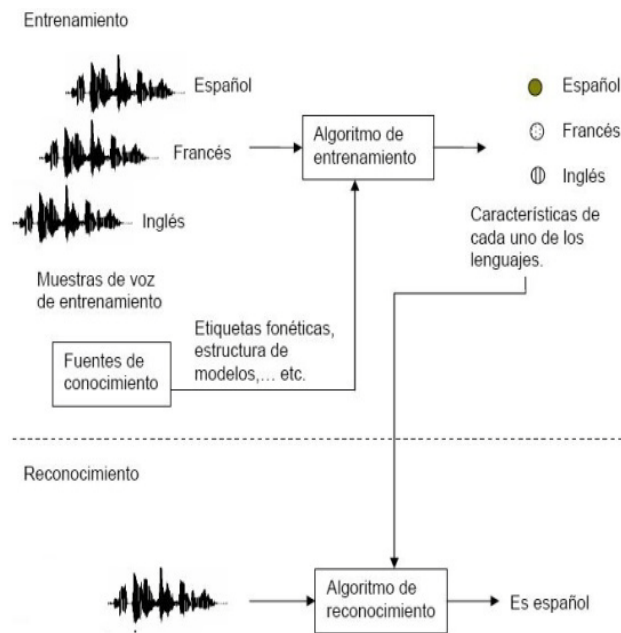


Fig. 1. Fases del proceso de identificación del lenguaje hablado

Los sistemas varían primeramente de acuerdo al método que utilizan para modelar los lenguajes. Por lo que existen diferentes métodos para extraer características de la señal de voz, esto ha incrementado la cantidad de conocimiento haciendo que los sistemas de identificación del lenguaje hablado sean cada vez más complejos. En el siguiente epígrafe se explican brevemente las fuentes de extracción de conocimiento que más se utilizan en estos sistemas [1].

2.1 Características del lenguaje hablado

Para representar un lenguaje en particular y diferenciarlo de otros es necesario entrenar un modelo de LID basado en discursos del lenguaje correspondiente. Por lo tanto un reto

importante para los sistemas LID basados en discursos es la incorporación efectiva a sus modelos de fuentes de conocimiento que sean discriminativas. Algunos sistemas usan sólo las expresiones digitalizadas de discurso y las correspondientes identidades conocidas de los lenguajes de las muestras de entrenamiento, mientras que los otros requieran información adicional como las transcripciones fonéticas y / o ortográficas, lo cual puede resultar muy caro. Durante la fase de reconocimiento de lenguaje, una muestra de audio nueva es comparada con cada uno de los modelos dependientes del lenguaje, y se selecciona el lenguaje que corresponda al modelo más cercano (por ejemplo usando la máxima verosimilitud). Para identificar un lenguaje existen varios niveles de conocimiento que pueden usarse, debiéndose investigar qué tipo de representación es la más apropiada para cada caso [10].

2.1.1 Rasgos derivados automáticamente de las señales del habla (rasgos acústicos)

Los lenguajes pueden ser identificados basándose en rasgos derivados automáticamente de señales del habla pertenecientes al propio lenguaje. Los sistemas que utilizan este tipo de información están motivados por la observación de que los diferentes lenguajes están constituidos por una variada colección de sonidos diferentes; por lo tanto, los vectores de características extraídos automáticamente sobre pequeños fragmentos de tiempo (segmentos) de las señales del habla pueden usarse para discriminar entre los lenguajes. Tales sistemas típicamente usan un proceso de múltiples pasos (que incluye módulos para eliminar el silencio de las muestras, reducir efectos del canal, etc.) para convertir la señal digitalizada de lenguaje en una representación del vector de características. Teniendo los rasgos y el conocimiento de cuales muestras de entrenamiento concuerdan con cada lenguaje, se construye un clasificador para cada idioma basado en los patrones dependientes del lenguaje de los vectores de características. Los métodos incluyen enfoques que modelan sólo la distribución estática de los rasgos acústicos aportados por el lenguaje [11] y enfoques que también utilizan los patrones de cambio de estos vectores de características con el paso del tiempo [7]. Siguiendo esta línea se ha investigado una variada colección de modelos computacionales que incluyen los modelos de mezclas gaussianas GMM [12], los modelos ocultos de Markov [12, 13], las redes neurales artificiales ANN [14], y las máquinas de vectores de soporte SVM[15]. Aunque estos sistemas basados en los rasgos acústicos no requieren que los datos de entrenamiento sean etiquetados con unidades lingüísticas explícitas como fonemas o palabras, tampoco actúan tan exactamente como los sistemas que usan conocimiento lingüístico; sin embargo, las mejoras en la exactitud pueden ser logradas integrando fuentes de conocimiento acústico y de bases lingüísticas [14, 16].

2.1.2 La información fonológica

La fonología es una rama de la lingüística que estudia los sistemas sonoros de los lenguajes humanos[17]. En un lenguaje dado, un fonema es una unidad simbólica en un nivel particular de representación; el fonema puede ser concebido como representante de una familia de pronunciaciones relacionadas que los parlantes de un lenguaje piensan que son categóricamente lo mismo. Mientras la noción de los fonemas ha sido un problema entre los lingüistas, ha

comprobado ser una abstracción útil para el procesamiento de voz. Ladefoged [18] supone que hay probablemente acerca de 600 consonantes diferentes provenientes de los lenguajes de mundo; las vocales y las distinciones suprasegmentales (como el tono) no son realmente tan numerosas.

Los símbolos fonéticos proveen una forma para transcribir los sonidos de lenguajes hablados. Hay varios alfabetos fonéticos; uno comúnmente usado por lingüistas es el Alfabeto Fonético Internacional (IPA) [19]. Este alfabeto, incluyendo un conjunto de modificadores diacríticos, fue establecido como un estándar por la Asociación Fonética Internacional para proveer un sistema figurativo preciso para transcribir los sonidos fonéticos de todos los lenguajes. El objetivo del IPA es proveer una representación para todo los fonemas expresados en todos los lenguajes humanos, tales símbolos separados sirven para dos sonidos, si y sólo si, existe un lenguaje para el cual estos dos sonidos son fonéticamente distinguibles.

Dada la disponibilidad de una representación fonética o fonémica para los sonidos de los lenguajes, hay diversas formas en las que esta información puede ayudar a diferenciar entre los lenguajes, incluyendo:

Inventario Fonémico: Cabe distinguir algunos lenguajes de otros basándose en la presencia de un fonema o pronunciación que aparece en un lenguaje pero no en el otro. Los inventarios de fonemas van desde un punto por debajo de 11 (para Rotokas, un lenguaje papú) [20] hasta por encima de cien (en ciertos lenguajes Khoisan del sur de África) [21]. Es probable que ninguno de los dos lenguajes compartan un inventario de fonemas exactamente igual. Además, aun si los dos lenguajes compartieran un conjunto común de fonemas, los fonemas mismos probablemente diferirán en sus patrones de frecuencia relativa [14].

Inventario de categorías fonéticas: Los patrones de categorías fonéticas amplias (ejemplo vocales, consonantes fricativas, explosivas, nasales, y líquidas) también ha sido utilizado para distinguir entre lenguajes en un intento para evitar la necesidad del reconocimiento fonético fino. Muthusamy [14] evaluó el uso de siete categorías fonéticas amplias (las vocales, la consonante fricativa, las paradas, el cierre o el silencio, sonido sonoro prevocálico, sonido sonoro intervocálico, y sonido sonoro postvocálico). Sin embargo, Hazen [16] encontró que, aun cuando los reconocedores anchos de categorías fonéticas tienden a ser más precisos que los reconocedores fonéticos tradicionales más minuciosos y finos, usando pistas anchas de inventarios de categorías los sistemas de identificación de lenguaje tienen menor exactitud.

Rasgos Fono Táctiles: Los rasgos fonotácticos se refieren a las disposiciones de las pronunciaciones fonéticas o los fonemas dentro de las palabras. Aun si dos lenguajes compartieran un inventario de fonemas común, es probable que difiriesen en sus rasgos fonotácticos. Los primeros en proponer el uso de estos rasgos fueron House y Neuburg [22], quienes creyeron que la identificación precisa del lenguaje podría ser lograda haciendo uso de las estadísticas de los acontecimientos lingüísticos en una sola palabra, en particular, las restricciones de secuencia específicas a la fonética del lenguaje. Ellos sugirieron usar una secuencia de categorías fonéticas amplias como una forma de obtener una extracción de características más fidedignas para los lenguajes, aunque se ha encontrado que esto resulta en sistemas de identificación de lenguajes menos precisos [16]. Una implementación usando este acercamiento típicamente implicaría varios pasos, donde la primera parte consiste en trazar un mapa de una sola palabra para una secuencia de etiquetas fonéticas (por ejemplo, la tokenización en sonidos fonéticos), que entonces se usaría para identificar el lenguaje basándose en los n-gramas observados.

Hay diversas variantes de este acercamiento para la identificación de lenguaje basada en señales de audio. Los LID basados en fonemas usan un solo reconocedor de sonidos fonéticos de lenguaje entrenado para algún lenguaje arbitrario con suficientes recursos (no necesariamente uno de los lenguas a identificar) para tokenizar la entrada de datos hablados en los sonidos fonéticos para ese lenguaje, y después usan los modelos probabilísticos n-gramas del lenguaje (uno para cada idioma deseado) para calcular la probabilidad de que la secuencia de símbolos haya sido producida por cada uno de los idiomas cuestionados, siendo seleccionado el lenguaje que provea la probabilidad más alta. El reconocimiento fonético paralelo seguido por el modelo de lenguaje (PPRLM) es similar excepto que usa reconocedores de sonidos fonéticos de varios lenguajes, conjuntamente con algún método para normalizar y combinar los resultados de los flujos paralelos [23].

Una tercera variante sería entrenar un reconocedor fonético con una base de datos fonética de amplia cobertura [18]. Para reducir el tiempo y el costo para desarrollar sistemas hablados en un lenguaje nuevo, los investigadores han estado investigando el desarrollo y el uso de un conjunto de sonidos fonéticos multilingües que representen los sonidos a través de los lenguajes que van a ser modelados. Schultz y Waibel [24] en su investigación sobre el reconocimiento del lenguaje hablado multilingüe definieron un conjunto de sonidos fonéticos que cubren 12 lenguajes. Ellos suponían que las representaciones articulatorias de los fonemas son tan similares a través de los lenguajes, que los fonemas pueden ser considerados como unidades que son independientes del lenguaje subyacente [25].

Hazen y Zue [16, 26] desarrollaron un sistema LID que se basó en 87 unidades fonéticas independientes del lenguaje que fueron obtenidos agrupando a mano aproximadamente 900 etiquetas de sonidos fonéticos encontradas entrenando las transcripciones. Los investigadores en el laboratorio de informática para mecánicos y las ciencias de ingeniería (LIMSI) también han estado investigando el uso de un conjunto de sonidos fonéticos universales [27, 28] y han tratado de identificar criterios acústicos objetivos para agrupar los sonidos fonéticos dependientes en lenguaje. Corredor-Ardoy [29], encontró que su sistema de LID usando un conjunto independiente en lenguaje (basado sobre tokens) se comportó tan bien como sus mejores métodos empleando sonidos fonéticos dependientes en lenguaje. Ma y Li [30] evaluaron el uso de un reconocedor universal de sonidos para transcribir expresiones en una secuencia de símbolos de sonido que actúan como un conjunto común de sonidos fonéticos para todos los lenguajes a ser identificados. Entonces usaron las estadísticas relacionadas con la co-ocurrencia de estos estándares de sonido en intervalos grandes para identificar el lenguaje de la expresión. A esta aproximación apodaron bolsa de sonidos.

Características articulatorias: El lenguaje puede ser caracterizado por flujos paralelos de características articulatorias que son usadas en la coordinación para producir una secuencia de fonemas. Estos rasgos pueden ser explotados para diferenciar un lenguaje de otro. Por ejemplo, el fonema /t/ puede realizarse con o sin aspiración, con un cierre dental o de los alvéolos, y con labios redondeados o no [16]. Kirchhoff y Parandekar [31] utilizaron un conjunto de categorías pseudo-articulatorias que fueron diseñadas para captar las características del proceso de producción del habla, que incluye: la manera de articular, el lugar consonántico de la articulación, el lugar vocálico de la articulación, el redondeando del labio, la posición de la lengua (trasera o delantera), la sonorización, y la nasalidad. Ellos desarrollaron un método alternativo para la identificación de lenguaje basado en audio que se basó en el uso de flujos

paralelos de estos eventos sub-fonéticos conjuntamente con el modelado de algunas dependencias estadísticas entre los flujos.

2.1.3 Estructuras de sílabas

Una sílaba es una unidad de pronunciación que es mayor que un solo sonido, compuesta por un pico de sonoridad (usualmente una vocal, excepto algunas veces una consonante sonora), limitado por sus puntos mínimos más cercanos (típicamente las consonantes) [17]. Los lenguajes pueden ser caracterizados por tipos comunes de sílaba, típicamente definidos en términos de secuencias de consonantes (C) y los vocales (V). El quebrantar estas secuencias hace que el proceso de silabeo sea a menudo difícil, e incluso discutible [32, 33]. Sin embargo, desde que los lenguajes generalmente permiten o prohíben ciertos tipos de estructuras de sílaba (ejemplo los lenguajes eslovacos a menudo tienen grupos de consonantes complicados en contraste a los lenguajes asiáticos), esta representación puede ayudar a discriminar entre los lenguajes. Por ejemplo CCCVC es un tipo válido de sílaba en ruso, pero no en la mayoría de otros lenguajes.

Zhu et al. [34] desarrollaron un sistema LID cuyo decodificador acústico produce un flujo de sílabas, que fueron usadas en modelos de lenguaje de n-gramas basados en sílabas (en vez de basados en sonidos fonéticos). El acento de los interlocutores extranjeros del idioma inglés se manifiesta diferentemente dada su localización dentro de la sílaba, hecho que se ha usado para mejorar identificación del acento en este lenguaje [35].

2.1.4 La información prosódica

La duración, el tono, y los patrones de énfasis en un lenguaje a menudo difieren de otro. Por ejemplo, los lenguajes diferentes tienen patrones bien definidos de entonación. En los lenguajes de énfasis, el tono es a menudo un correlativo de acentuación usado para señalar la prominencia de la sílaba en las palabras (y en los lenguajes de tono acentuado, como el sueco, el correlativo es primario); mientras que, en los lenguajes de tono, un cambio en el significado de una palabra es señalado por el tono en las sílabas u otras unidades que producen tono (el Chino-Mandarín o el Tailandés). Para los lenguajes de énfasis, los patrones de énfasis pueden proveer una pista importante para discriminar entre dos lenguajes. Algunos de los patrones de énfasis son el énfasis inicial (el húngaro), el énfasis penúltimo (el polaco o español), el énfasis final (francés o turco), y el énfasis mixto (ruso o griego). Las pistas prosódicas de duración también pueden ser potencialmente útiles. Por ejemplo, algunos lenguajes (como finés) hacen la distinción por vocales y / o las consonantes, breves y largas.

Una tipología notablemente útil investigada por lingüistas y psicólogos implica el ritmo de los lenguajes, existiendo diferencias entre lenguajes de los que se midió el tiempo de énfasis (en el que las sílabas acentuadas resultaron ser más largas que las sílabas sin acentuar, si no todo sería igual, ejemplo, el inglés), el tiempo de cada sílaba (cada sílaba tiene duración de tiempo comparable, ejemplo, el francés), y el tiempo de demora culpable (cada demora culpable tiene duración esencialmente constante, ejemplo, el japonés) [36]. Aunque esta clasificación permanece en constante discusión [37], el modelado rítmico fue investigado por Rouas et al. [38, 39] para la identificación de lenguaje. Su modelo de ritmo pudo discriminar acertadamente con

claridad entre los lenguajes a partir de la lectura de un texto fijo, pero no tan así a partir de un discurso espontáneo [38]. Ciertamente, extraer pistas prosódicas fidedignas a partir de un discurso espontáneo constituye un reto debido a su variabilidad.

En los sistemas de identificación de lenguaje que utilizan características prosódicas, estas características generalmente son combinadas con otras fuentes de conocimiento para lograr una exactitud razonable. Muthusamy [14] fue capaz de incorporar la variación del tono, la duración, y las características de proporción de sílaba en su modelo para LID. Hazen y Zue [16, 26] integraron información de duración y de tono en su modelo para LID, resultando ser más preciso el modelo de duración que el del tono. Tong et al. [40] integraron exitosamente características prosódicas (duración y tono) del espectro, fonotácticas, y características de bolsa de sonidos, donde la variación del tono y duración de los fonemas fueron especialmente útiles para los pequeños segmentos de discurso.

2.1.5 Información léxica

Cada lenguaje tiene su propio vocabulario, lo cual debería ayudar en identificar un lenguaje más confiadamente. Para las expresiones de entrada, esto requeriría la disponibilidad de un sistema de reconocimiento del lenguaje hablado por cada uno de los lenguajes candidatos, junto con el entrenamiento requerido, la afinación, y los materiales de evaluación necesarios para asegurar que el modelo del lenguaje es adecuado. Schultz et al. [41, 42] desarrollaron un sistema LID para cuatro lenguajes basándose en el amplio léxico del reconocimiento del lenguaje hablado continuo (LVCSR). Ellos encontraron que los sistemas basados en palabras con modelos trigramas de palabras funcionan significativamente mejor que los sistemas basados en fonemas con modelos trigramas de sonidos fonético para la identificación de de cuatro lenguajes, sugiriendo que el nivel léxico provee habilidad de discriminar entre los lenguajes, aún cuando las tasas de errores de palabra son medianamente altas. Matrouf et al. [43] encontraron que incorporando información léxica con un enfoque basado en fonemas se produjeron disminuciones relativas de error del 15–30%, y este alcance incremental del léxico para un lenguaje tuvo un efecto positivo sobre el funcionamiento del sistema. Hieronymus and Kadambe [44] construyeron un sistema LID basado en LVCSR para cinco lenguajes (chino, inglés, alemán, japonés, mandarín, y español), y obtuvieron un 81% y 88% de identificación correctas dados 10 y 50 segundos de expresiones, respectivamente, sin usar medidas de confianza y 93% y 98% de identificaciones correctas con medidas de confianza. Aunque cada lenguaje obviamente tiene su vocabulario que le permite a los sistemas de identificación de lenguajes discriminar entre un conjunto de candidatos de lenguajes más eficazmente, para los lenguajes hablados, esta información es generalmente bastante cara para obtener y usar.

2.1.6 Información morfológica

La morfología es una rama de la gramática que investiga la estructura de las palabras [17]. El campo de la morfología está dividido en dos sub-campos: La morfología aglutinativa, que investiga afijos que señalan las relaciones gramaticales que no cambian la categoría gramatical de una palabra (afijos que marcan tiempo, número, y eventos) y la morfología derivativa, la cual

se centra en la formación de palabras requiriendo de afijos, algo semejante a la terminación “ción”, que puede ser usado para crear una nueva forma de palabra con una categoría de gramática posiblemente diferente, como el sustantivo *modificación* deducido del verbo *modificar*. Como que los lenguajes forman palabras siguiendo diferentes formas, la morfología podría proveer una pista excelente para identificación automática de los mismos. Por ejemplo, uno puede emplear sufijos comunes para discriminar entre algunas de las lenguas romance (-ment en francés, -miento en español, -mento en Portugese, y -mente en italiano). Aunque en el lenguaje la morfología de una palabra es poco notoria en parte por la información fonotáctica, considerando el conocimiento morfológico de los lenguajes a identificar, un sistema LID podría enfocarse sobre porciones específicas de palabras para discriminar entre dos lenguajes.

2.1.7 Información sintáctica (Syntaxis)

Los lenguajes y los dialectos también difieren en las formas en que las palabras son arregladas para crear una frase. Difieren en la presencia o la ausencia de palabras en diferentes partes de la oración, así como también en las formas en que las palabras son acentuadas para varios tipos de roles en una frase. En conjunto con otros tipos de información (los acentos), los errores en el uso gramatical podrían proveer una pista útil para identificar la lengua materna de alguien hablando un segundo idioma. Por ejemplo, alguien que aprendió mandarín como lengua materna tendería a cometer determinados errores de armonía en inglés o alemán (la supresión, la sustitución, y la inserción).

Los lenguajes a menudo también difieren el uno del otro en la orden de las palabra dentro de una oración sujeto (S), verbo (V), y objeto (O). Por ejemplo, se considera que el inglés es un lenguaje SVO porque el sujeto típicamente aparece delante del verbo, lo cual ocurre antes del objeto; mientras que, el japonés es un lenguaje SOV. Aun si dos lenguajes tienen la misma forma de organizar la información, es probable que las palabras apareciesen en contextos de frases muy diferentes a través de los lenguajes. Aunque las palabras pueden ser suficientes como para discriminar entre los lenguajes, la syntaxis podría jugar un papel especialmente importante para identificar los diferentes dialectos de un lenguaje.

Los sistemas de la identificación de lenguajes y dialectos que se basan en LVCSR podrían utilizar un modelo acústico, un diccionario, y un modelo de lenguaje para cada idioma o cada dialecto a identificar. Los modelos de lenguaje utilizan estadísticas de la co-ocurrencia de palabras que captan algunos aspectos de la estructura sintáctica del lenguaje candidato. Estos sistemas podrían ser expandidos para utilizar la syntaxis más directamente usando modelos estructurados de lenguaje [45, 46].

Las investigaciones en la identificación automática de lenguaje sugieren que se necesita mucho conocimiento para examinar a fondo una muestra de audio y tomar una decisión acerca cual lenguaje concuerda con ella, con la mayor la exactitud; por lo tanto, la integración de conocimientos es importante. Sin embargo, hay una relación de correspondencia entre la eficacia y la eficiencia, y además, está la necesidad de poseer recursos para sustentar el conocimiento introducido en el sistema automático. En este sentido se ha llegado a un balance ingenieril; usando los recursos fidedignos que pueden ser obtenidos de la forma más simple para el conjunto de lenguajes que va a ser discriminado.

Es importante resaltar que la longitud de una muestra de audio (la cantidad de discurso disponible) afectará las fuentes de conocimiento que pueden utilizarse confiadamente determinando su lenguaje. Mientras más pequeñas sean las muestras usadas, menos probable es que un pedazo significativo de conocimiento a alto nivel esté disponible para discriminar un lenguaje particular de los demás.

3 Tendencias prevalecientes en la Identificación automática del lenguaje hablado

La identificación automática del lenguaje hablado (LID) es el proceso de determinar el idioma hablado en una expresión tomada como muestra.

Una fase típica del reconocimiento de patrones es la evaluación de la probabilidad subsecuente $P(\lambda_l | X)$, del l -ésimo lenguaje que va a ser considerado en el modelo λ_l dada una expresión conocida X . Para tomar la decisión, el idioma que produce la máxima $P(\lambda_l | X)$ es generalmente identificado como el idioma en cuestión. Muchos algoritmos desarrollados en reconocimiento automático del habla y del locutor pueden ser adaptados y extendidos al reconocimiento del idioma [47, 48]. Las recientes avances en el modelado acústico y del lenguaje [49] también han contribuido al progreso del reconocimiento del idioma.

3.1 Sistemas con reconocimiento fonético

La figura 2 muestra los componentes básicos para la identificación del lenguaje hablado basado en la representación fonética de la señal de voz. Este tipo de sistemas utiliza dos tipos de información: la fonética y la fonotáctica (reglas que rigen la distribución de los sonidos fonéticos dentro de un lenguaje), y consta de dos pasos básicos:

Reconocimiento fonético: En este primer paso es donde tiene lugar *el procesamiento acústico de la señal de voz*, proceso que consiste en la transformación de la señal de voz continua (ondas acústicas) en una secuencia de eventos discretos a partir de los cuales será posible identificar los fonemas dentro de la señal estableciendo un modelo acústico para cada uno de ellos. Ya en una segunda fase es necesario hacer un *Alineamiento de fonemas*, considerado como un paso fundamental para poder establecer la captura de las características estructurales de niveles altos. Para ello existen diferentes métodos tales como, los *clusters* de fonemas, mapeo de fonemas, mezcla de lenguajes, etc. Para finalizar el reconocimiento fonético se pasa entonces la *Extracción de características estructurales*, tales como, palabras, subpalabras, sílabas, etc. En los métodos basados en la dependencia del lenguaje se obtienen secuencias de fonemas válidas de cada lenguaje o categorías de fonemas que se corresponden directamente a diferentes sonidos en el lenguaje de cada hablante.

Modelado del lenguaje: Es a partir de la secuencia de fonemas que propiamente se crea un modelo para cada lenguaje. Este no es más que un listado de todas las posibles combinaciones de dos, tres o n fonemas con sus respectivas probabilidades (las características fonotácticas del idioma). Los modelos pueden ser entrenados usando modelos ocultos de Markov (HMM), redes neuronales o mezclas gaussianas, entre otros.

A la hora de identificar una muestra desconocida pues se parte de la secuencia de fonemas que fue propiamente extraídos y se comparan las proporciones de los mismos en contra los modelos de los lenguajes a identificar previamente creados durante el entrenamiento del sistema.

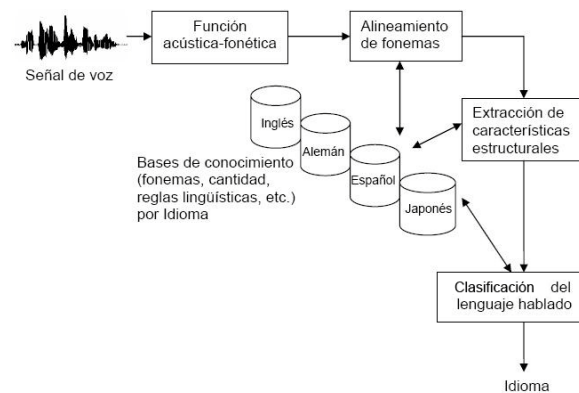


Fig. 2. Diagrama de diseño de los sistemas LID que utilizan representación fonética

Es importante destacar que la construcción de los modelos acústicos es necesario contar con un gran número de grabaciones, que cubran todos los fonemas del idioma así como sus alófonos (variantes de los fonemas dado su contexto, el acento regional, etc.). Ahora bien, dichas grabaciones deben estar etiquetadas manualmente por expertos a nivel fonético, es decir, cada grabación debe ser escuchada por el experto para determinar las fronteras entre los fonemas. Por otra parte para la creación de los modelos de lenguaje son necesarios grandes corpus de texto (y/o transcripciones ortográficas de grabaciones) que permitan la estimación de las probabilidades de co-ocurrencia de estas unidades lingüísticas para cada uno de los idiomas. Tanto la segmentación manual de la señal como los procesos de transcripciones resultan tareas muy costosas, que se traducen en las principales desventajas de estos sistemas. [1]

Zissman y Singer [50], y posteriormente Torres-Carrasquillo et al [7], trataron de mejorar un poco esta situación incluyendo la categoría “token” en la fase de reconocimiento fonético. Un token es una unidad lingüística que está por encima de los fonemas. Partiendo del mismo concepto, un token sería una secuencia de unidades fonéticas (fonemas). Es decir, cada sonido fonético y sus contextos circundantes se combinan para formar n-gramas de fonemas, resultando en varios tokens. Intuitivamente, ellos representan las dependencias fonotácticas existentes entre las unidades fonéticas individuales. Los valores resultantes de estas secuencias de fonemas simbolizados fueron utilizados en la identificación de los lenguajes hablados obteniendo mejoras significativas en los resultados de los sistemas LID puesto que concentraban mejor la información ganando en cuanto al nivel de profundidad con que debían trabajar los especialistas a la hora etiquetar las señales de entrenamiento.

En las próximas sesiones se exponen brevemente los detalles correspondientes a los sistemas típicos que se han desarrollado bajo este enfoque. Se hace notar que en varias ocasiones en el

texto se refieren a los procesos de reconocimiento fonético con el término de tokenización de la señal.

3.1.1 Modelación del lenguaje a partir de reconocimiento fonético. PRLM

Este es un sistema de identificación de lenguaje constituido por un reconocedor de sonidos fonéticos específicos a un lenguaje (generalmente el que se pretende identificar), seguido por un analizador n-gramas, como se muestra en la figura 3 [50-52] .



Fig. 3. Diagrama de bloques de un sistema PRLM

En este sistema, los mensajes de entrenamiento de cada lenguaje son procesados por un reconocedor fonético a fin de obtener una secuencia de caracteres resultante asociada con cada uno de los mensajes de entrenamiento. Una de las opciones para modelar los fonemas (o tokens) sería considerar la información secuencial haciendo uso de los modelos ocultos de Markov (HMM) entrenados sobre expresiones no etiquetadas. Aunque recientemente se han obtenido algunas mejoras [53], no se ha logrado obtener mejores resultados con esta técnica que con la clasificación estática [12, 54]. Una buena elección sería modelar las unidades fonéticas usando los Modelos de Mezclas Gaussianas (GMM), lo que permitiría generar la secuencia de caracteres desde una densidad específica del lenguaje. Esta aproximación es puramente basada en ventanas sin que se utilice la información segmental y es exactamente la misma formulación que se hace en la identificación de locutores basados en GMM con independencia del texto [55].

Una vez obtenida la secuencia de caracteres resultante asociada con cada uno de los mensajes de entrenamiento se analiza y se estima un modelo de distribución de probabilidad de n-gramas para cada lenguaje L . Aquí el modelo del lenguaje λ_L puede ser cualquier caracterización lógica de los vectores de rasgos. Las Máquinas de Soporte Vectorial SVM [56] han estado siendo usadas para diseñar sistemas de reconocimiento de locutores y de idiomas basados en GMM ambos con altos rendimientos [8, 9, 57].

Durante el proceso de reconocimiento, el mensaje experimental es sometido al reconocimiento fonético y se calcula la probabilidad de que su secuencia de caracteres haya sido producida por cada uno de los lenguajes en cuestión. El modelo de n-gramas que dé como resultado la probabilidad más alta es identificado, y el lenguaje de ese modelo será seleccionado como el lenguaje del mensaje.

Para obtener modelos de los fonemas (o tokens) y de idiomas con alta exactitud, para cada idioma individual sobre la consideración de que esto no siempre es posible, se puede considerar entrenar un conjunto de modelos de unidades fonéticas individual que cubran todos los idiomas y usarlo para decodificar todas las expresiones habladas.

Resulta interesante hacer notar que dichos modelos son usados solo para decodificar, y no para calcular las puntuaciones a la hora de determinar el idioma identificado. Además se puede expandir la colección de modelos de fonemas (o de tokens) para incluir varios conjuntos de

modelos, cada uno entrenado por ejemplos hablados de un subconjunto de lenguajes. El conjunto de F modelos de unidades fonéticas $\{\lambda_1^{AM}, \dots, \lambda_f^{AM}, \dots, \lambda_F^{AM}\}$ serían todos usados para decodificar una expresión hablada dada, dando como resultado un conjunto de F secuencias de fonemas (o tokens). Además de esto, para el i -ésimo idioma, podemos entrenar F conjuntos de modelos fonéticos del lenguaje, $\{\lambda_{i,1}^{AM}, \dots, \lambda_{i,f}^{AM}, \dots, \lambda_{i,F}^{AM}\}$, para calcular las F puntuaciones de un idioma específico, $\{P(q_f | \lambda_{i,f}^{AM}), f = 1, \dots, F\}$, las que pueden ser combinadas para tomar decisiones en la clasificación de los lenguajes. Esta es la idea básica detrás del PRLM paralelo (P-PRLM), [23] el cual es, en cuanto a rapidez, la más exitosa aproximación para la identificación de lenguajes LID [10].

3.1.2 Modelación del lenguaje a partir de reconocimiento fonético en paralelo. P-PRLM

Aunque PRLM es una manera efectiva de identificar el lenguaje de un mensaje de habla, se conoce que el lenguaje que se utiliza para entrenar el reconocedor fonético no siempre abarca todos los sonidos que ocurren en los lenguajes a identificar. Por esta razón, se necesita buscar una forma para incorporar fonemas (o tokens) de más que un lenguaje en un sistema análogo al PRLM. En este sentido se ha planteado que una posible solución sería entrenar el reconocedor fonético con expresiones de varios lenguajes [58]. Alternativamente, este acercamiento plantea simplemente correr múltiples sistemas PRLM de forma paralela usando un conjunto de reconocedores fonéticos simples cada uno entrenado para un lenguaje diferente [50, 59]. Una desventaja de este sistema es que necesita etiquetar las señales de entrenamiento para más de un lenguaje, aunque no necesitan estarlo para todos los lenguajes a reconocer, o incluso, para ninguno de los lenguajes a reconocer, solo para aquellos, con los que se entrenan los reconocedores fonéticos.

La figura 4 muestra un ejemplo de un sistema PRLM en paralelo. Aquí se tiene acceso a las expresiones etiquetados para los idiomas japonés, inglés, y español, pero la tarea principal es realizar la clasificación del lenguaje de mensajes en Farsi, francés, y Tamil. Para ello, primero se entrenan tres sistemas separados PRLM: Uno con un reconocedor fonético para inglés, otro con un reconocedor fonético para japonés, y el último entrenado para español. Este sistema paralelo PRLM tendría un total de nueve modelos de lenguaje de n -gramas, uno por cada lenguaje a ser identificado (Farsi, francés, y Tamil) para cada reconocedor fonético (japonés, inglés, español).

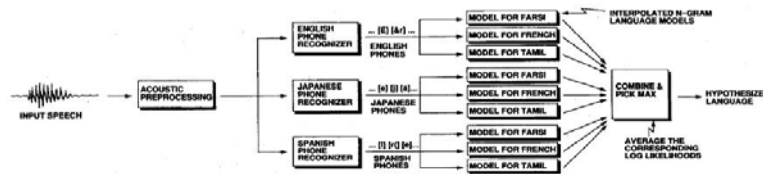


Fig. 4. Diagrama de bloques de un sistema P-PRLM

Durante el reconocimiento, el mensaje es procesado por los tres sistemas PRLM, y sus salidas son promediadas en un dominio logarítmico (multiplicando en el dominio lineal, como si cada sistema PRLM estuviera operando independientemente) para calcular las puntuaciones globales de probabilidad logarítmicas del lenguaje. Es interesante notar que este acercamiento se extiende fácilmente hasta cualquier número de sistemas paralelos PRLM y que su única limitación es el número de lenguajes para los cuales tienen que estar etiquetadas las señales de entrenamiento. Los parámetros del reconocedor fonético (número de estados, número de mezclas Gaussianas, etc.) son los mismos que se utilizan en PRLM[23].

3.1.3 Reconocimiento fonético en paralelo (PPR)

Para la implementación de los sistemas anteriormente vistos (PRLM y P-PRLM) se necesita disponer de señales de entrenamiento etiquetadas para cada uno de los lenguajes a identificar, lo cual no siempre es posible. La disponibilidad de tal recurso amplía el alcance de las posibilidades de los métodos de identificación de lenguaje, por ejemplo se vuelve fácil de entrenar y usar modelos acústicos/fonotácticos integrados. Hallando un reconocedor fonético que use las restricciones fonotácticas específicas de cada lenguaje durante el proceso de decodificación en vez de aplicar estas restricciones después de que el reconocimiento fonético es completado (como sucede en PRLM y PPRLM), las secuencias fonéticas más probables identificadas durante el reconocimiento son óptimas en relación a alguna combinación entre las acústicas y las fonotácticas por separado. La unión de estas probabilidades resultan muy apropiadas para la identificación del lenguaje, puesto que PPR también aprovecha las diferencias fonotácticas entre los lenguajes [60, 61].

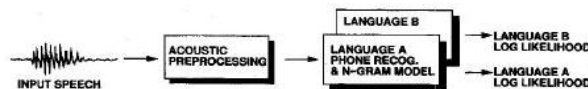


Fig. 5. Diagrama de bloques de un sistema PPR

Como se puede apreciar en la figura 5 los sistemas basados en este acercamiento tienen la misma configuración que los sistemas que usan reconocedores fonéticos de un solo lenguaje PRLM (usando HMM o GMM para modelar los fonemas) pero con la excepción de que el modelo del lenguaje forma parte integral del reconocedor (que en los sistemas PRLM esto se hace después de procesar). Otra diferencia es que ambos pueden hacer modelos fonéticos dependientes del contexto, solo que PRLM solo puede usar para ello mono fonemas, mientras que PPR usa los mono fonemas de cada lenguaje además de los 100 fonemas dependientes del contexto que ocurren comúnmente con más éxito. Esta estrategia estaba motivada por los resultados de los primeros experimentos que mostraban que los sistemas LID_PPR con fonemas dependientes del contexto mejoraron su funcionamiento, pero esto no tuvo ningún efecto en el funcionamiento de los LID_PRLM.

Como deficiencia tiene que requiere señales etiquetadas para cada lenguaje que vaya a reconocer, lo que lo hace mucho más difícil de implementar que el resto de los vistos hasta ahora.

3.2 Sistemas sin reconocimiento fonético

Los sistemas que se basan en este enfoque parten de eliminar los módulos que tienen que ver con la fase de representación fonética de la señal tal como muestra la figura 6, es decir, no se toma en cuenta la información lingüística del idioma a identificar. Por supuesto, la etapa de extracción de rasgos acústicos de la señal de voz es diferente a los sistemas anteriores puesto que se aprovechan otras características del lenguaje, que están presentes en la señal, tales como la prosodia, el ritmo, la entonación, las características espectrales, entre otras [1].

Los sistemas LID diseñados bajo este enfoque están basados en tramitar estos rasgos de bajo nivel. Los métodos que se utilizan para ello, intentan extraer medidas del espectro de la señal de habla de corto termino sobre un análisis de trama fija y entonces modelar las características de estas medidas, o rasgos, por cada lenguaje a reconocer. Las técnicas de clasificación que han presentado buenos resultados en esta tarea son las de corte generativo GMM, y discriminativo SVM. En los próximos epígrafes se describen algunos de los rasgos que mejores resultados han dado y las principales técnicas de clasificación que se utilizan para el reconocimiento y modelación de los lenguajes.

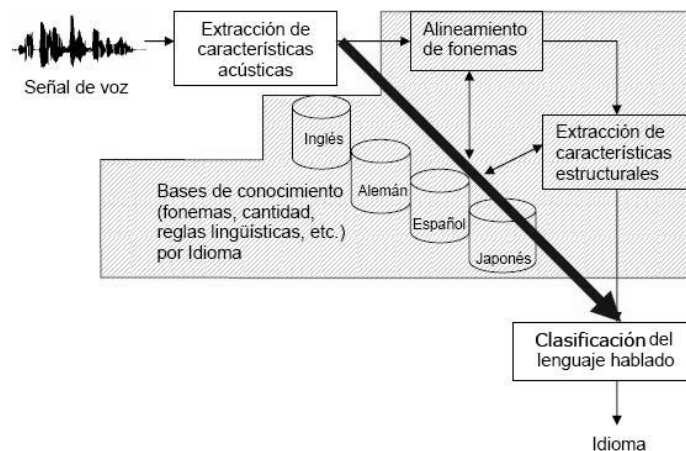


Fig. 6. Diagrama de representación de los sistemas LID que no utilizan representación fonética

3.2.1 Caracterización de los rasgos acústicos del lenguaje

Existen diferentes formas de extraer las características de la señal del habla digitalizada, teniendo en cuenta qué característica se desea obtener es que se utiliza una u otra forma de análisis. La búsqueda de rasgos acústicos que resulten discriminativos y robustos para la identificación del lenguaje ha resultado una tarea importante para el desarrollo de esta

tecnología. Una gran parte de las investigaciones realizadas en este sentido estuvieron dadas a probar aquellos rasgos que mostraron resultados muy provechosos en la discriminación de los locutores con el objetivo de ver si ocurría lo mismo en los lenguajes. Pero los resultados no fueron buenos. De hecho, utilizando este tipo de caracterización aun no se han podido superar los resultados expuestos por los sistemas que operan basados en el reconocimiento fonético. Los mejores resultados han estado orientados a la fusión (concatenación) entre varios tipos de rasgos, creando vectores de características cada vez más complejos.

A continuación se explica brevemente algunos de los rasgos que más se han utilizado y las tendencias de los mismos.

Coefficientes Cepstrales de Frecuencia Mel (MFCC) y coeficientes Perceptuales de Predicción Lineal (PLP)

Las características Cepstrales, representan típicamente las propiedades de magnitud de espectro de una señal de habla, por lo que son ampliamente usadas en el procesamiento de voz. Para lograr un rendimiento alto en los sistemas es importante escoger rasgos efectivos. Los coeficientes cepstrales de frecuencia Mel (MFCC) y una de sus alternativas, coeficientes perceptuales de predicción lineal (PLP), fueron por mucho tiempo los dos rasgos cepstrales más populares [62].

Los MFCC son utilizados para la representación del habla basándose en la percepción auditiva humana. Para su obtención se parte de la transformada de Fourier, teniendo en consideración que las bandas de frecuencia se sitúan logarítmicamente, según la escala de Mel (escala logarítmica desarrollada para representar mejor el patrón de percepción del oído humano). Davis y Mermelstein [63] fueron los primeros en mostrar una representación de los MFCC para el reconocimiento del habla. La misma representación fue adoptada por los investigadores para el reconocimiento del locutor, demostrando ser altamente exitosa en esta área [64-66].

Por su parte el PLP constituye una aproximación razonable al habla producida por la excitación del tracto vocal causada por el conjunto de pulsos glotales. El método se fundamenta en establecer un modelo de filtro del tipo todo polo, para la fuente de sonido. Los coeficientes LPC brindan una información combinada sobre la frecuencia, ancho de banda de los formantes y la onda glotal, siendo bastante robustos al ruido. Al igual que los MFCC fueron ampliamente usados en los sistemas de identificación del locutor obteniendo muy buenos resultados [67].

En análisis y experimentos salen a la vista las diferencias entre MFCC y PLP, no influyendo estas en el funcionamiento de los sistemas de reconocimiento del lenguaje hablado [68, 69]. Puesto que los MFCC y el proceso de extracción PLP son similares, la información inferida de ellos será similar también. Esa es la razón por la que la combinación de estos dos rasgos no produce una mejora valiosa.

Rasgos Cepstrales Delta Desplazados (SDC)

Aun cuando se conoce que los rasgos cepstrales son una mejor aproximación matemática a los rasgos del habla que el modelo PLP, es importante señalar que estos no contienen información sobre la variación en el tiempo de las características propias de la señal de voz. Los rasgos que contienen la información de la señal teniendo en cuenta su variación en el tiempo son denominados rasgos dinámicos. Ellos se obtienen a partir de los rasgos cepstrales utilizando herramientas matemáticas que reflejen la variabilidad en el tiempo. Para obtener un vector de rasgos dinámicos se calculan las primeras derivadas en el tiempo de cada uno de los coeficientes que conforman el vector de rasgos espectrales y se le anexan a dicho vector. A estos nuevos

rasgos obtenidos a partir de las derivadas se les denomina rasgos delta (Δ), ellos portan la información dinámica de los rasgos espectrales [70]. Comúnmente se estiman también las derivadas en el tiempo de los rasgos delta, conocidas como rasgos delta-delta ($\Delta \Delta$), y se anexan al vector de rasgos. Tal es el caso de los Rasgos Cepstrales Delta Desplazados (SDC).

Los rasgos SDC se utilizan frecuentemente en la identificación de idioma y dialecto obteniéndose buenos resultados. Ellos contienen información dinámica de la señal de voz, por lo que aportan gran cantidad de elementos distintivos del lenguaje.

La fuente de los rasgos SDC es una matriz de rasgos delta (Δ) de dimensión (N, T), los que a su vez se conformaron a partir de una matriz de rasgos cepstrales. Para su obtención se definen los siguientes 4 parámetros (N, D, P, k): [2, 71]

N: dimensión de los vectores de rasgos cepstrales (MFCC).

D: cantidad de rasgos a escoger para calcular el rasgo delta (Δ).

P: número de rasgos a desplazarse para calcular el próximo rasgo SDC.

k: número de vectores delta (Δ) que se concatenan para formar el vector final SDC.

Para obtener una matriz de rasgos SDC se parte de la obtención de una matriz de rasgos delta (Δ). Luego se toman k vectores de N rasgos delta (Δ) para un valor de t, espaciados P tramas, y se concatenan en columna. Es decir, el vector de rasgos SDC final queda formado por la concatenación de $i = 0$ a $k-1$ de todos los $\Delta c(t + iP)$:

$$\Delta c(t + iP) = c(t + iP + d) - c(t + iP - d)$$

Una matriz de rasgos SDC (N, d, P, k) está formada por kN vectores en columnas y t filas. Donde N es la cantidad de vectores de rasgos fuente, k es el número de rasgos delta (Δ) que se toma para formar el SDC y t la cantidad de instantes de tiempo en que se toman muestras. Los valores que tienen D y P intervienen en el conjunto de componentes del rasgo fuente que influyen en la formación de un elemento de la matriz de los rasgos SDC.

Otra forma de calcular los elementos del vector de rasgos SDC y que arroja mejores resultados para identificación del lenguaje es utilizar el mismo ajuste polinomial que se utilizó para calcular los delta (Δ) pero incluyendo el desplazamiento en tiempo iP entre bloques: [72]

$$\Delta c(t + iP) = \frac{\sum_{d=-D}^D dc(t + iP + d)}{\sum_{d=-D}^D d^2}$$

Es importante destacar que con esta técnica se tiene en cuenta tanto la información estática como la dinámica, puesto que se agrupa una mayor cantidad de elementos inherentes de la señal. Su principal ventaja radica en que al incorporar en cada trama de tiempo el comportamiento dinámico de tramas posteriores, se incrementa el valor identificativo del rasgo. Pues se recoge más tiempo de señal en la evaluación de la dinámica, reflejando las transiciones entre fonemas e incluso sílabas, lo que conduce a una mejora en la efectividad de la verificación del lenguaje.

Rasgos Prosódicos

La prosodia también desempeña un papel crucial en la percepción humana del habla. La información contenida en rasgos prosódicos es en parte diferente a la información contenida en las características cepstrales. Por consiguiente, cada vez más los investigadores del área de

reconocimiento del lenguaje hablado muestran interés en las características prosódicas ya que ellas encierran lo que los especialistas lingüistas llaman *información suprasegmental*. Varios investigadores han presentado sus enfoques al utilizar esta información en identificación de lenguaje. Estos métodos incluyen el contorno de tono en aprendizaje sin supervisión con GMM, siguiendo por la fusión con otros resultados de rasgos diferentes; [73] y el aprendizaje sin supervisión con HMM [74]. Sin embargo, la mayoría de los resultados de esta investigación salieron a la vista que la información prosódica tiene sólo buenas consecuencias para la identificación del tono en los lenguajes.

Generalmente, la prosodia quiere decir "la estructura que organiza el sonido" [75]. El tono, la sonoridad, y las estructuras de ritmo (el tiempo) son los componentes principales de prosodia. Típicamente, estos rasgos incluyen tono, intensidad, y la duración normalizada de sílabas, características que pueden ser extraídas basándose en técnicas de ventaneo. También se utilizan técnicas de eliminación de silencio puesto que esto no aporta ninguna información interesante sino sólo ruido, lo que es una cuestión fundamental a tener en cuenta al utilizar rasgos prosódicos.

Ahora bien, recientemente se han desarrollado varias técnicas para capturar la información suprasegmental existente en las señales de habla. Cummins et al [76] basan el procesamiento acústico solamente en el uso de la frecuencia fundamental. Samouelian [77] amplía el conjunto de características al usar las frecuencias secundarias, con el uso de 12 MFCC (cantidad de coeficientes comúnmente usada para el procesamiento acústico en el tratamiento de voz [78]) calculando los cambios o deltas en ventanas adyacentes. Se conoce que los cambios temporales en el espectro juegan un papel muy importante en la percepción humana y que una de las maneras de capturar esta información es el uso de estos cambios o deltas en los coeficientes cepstrales. Con ellos es posible describir el cambio de cada coeficiente en el tiempo, capturando así, los cambios de la señal de voz, cambios posiblemente entre los fonemas, las sílabas y quizás hasta palabras, buscando de esta forma capturar parte de la información suprasegmental. Este hecho demuestra la utilidad de los MFCC tanto en la identificación del lenguaje hablado con representación fonética, como en la identificación sin reconocimiento fonético.

Ana [1] en su investigación doctoral propuso dos nuevos métodos de extracción de características específicos para la identificación del lenguaje hablado, basados en las características suprasegmentales, distintivas entre los idiomas.

El primero de ellos plantea capturar los cambios temporales en el espectro de la señal de voz por medio del uso de 16 MFCC, capturando además de la frecuencia fundamental F_0 , las frecuencias secundarias. Se propuso además el uso de los deltas Δ_1 , Δ_2 y Δ_3 , que pretenden capturar el cambio de un coeficiente cepstral entre una ventana y su adyacente. Con estos deltas, se busca capturar las diferencias que hay entre los fonemas y posiblemente entre sílabas. Como resultado se obtuvo un nuevo método de caracterización de la señal de voz, con 192 atributos. Otra observación que se hace es que con muestras pequeñas de señal de voz se obtuvieron mejores resultados, al contrario de lo que se esperaba puesto que los deltas tienden a estabilizarse para muestras de señal de voz más grandes. Probablemente por la mayor aparición de pausas y silencios.

El segundo método, tuvo como idea central obtener una nueva caracterización orientada a las bajas frecuencias. Partiendo de que la frecuencia fundamental es la más baja de todas y fue el parámetro más utilizado por los métodos del estado del arte para representar la prosodia. Por lo tanto se asume que en las frecuencias bajas hay información relevante para la identificación del

lenguaje hablado; concretamente las características suprasegmentales que usamos al hablar. Por esta razón se propone el uso de la transformada wavelet para el procesamiento de la señal de voz; ya que esta tiene una muy buena resolución en las bajas frecuencias haciendo una separación entre estas y las altas frecuencias. Este método es completamente diferente a los usados anteriormente basados en la transformada de Fourier.

La principal diferencia contra el primer método propuesto es que el ventaneo de la transformada de Fourier de tiempo corto (STFT) es fijo en contra del ventaneo de la wavelet que es variable. Teniendo en cuenta que el ancho de ventana constituye un parámetro de gran importancia ya que a través de éste podemos establecer el grado de resolución tanto de tiempo como de frecuencia. Se conoce que para tener una buena resolución de las altas frecuencias se necesitan ventanas pequeñas y para una buena resolución de las bajas frecuencias se necesitan ventanas grandes lo que resulta importante, ya que se desea capturar la información suprasegmental del habla.

Otra diferencia a considerar es que los dos métodos son sensibles al tamaño de muestra de la señal de voz, para el caso de la extracción por medio de wavelet se necesitan muestras grandes, alrededor de 50 segundos para obtener muy buenos resultados, no siendo así para el caso de la extracción de características suprasegmentales por medio de MFCC. Cabe mencionar que las wavelet han sido utilizadas en el reconocimiento del habla, [79, 80] pero hasta ahora no se había reportado ningún trabajo aprovechando el uso de wavelet en la identificación de idiomas.

Características Modificadas de Retraso del Grupo (MODGDF)

Como se puede apreciar las investigaciones en los últimos años han estado orientadas a trabajar con las características de magnitud del espectro (señal en forma discreta), dejando en evidencia que los intentos para utilizar el espectro de fase para obtener características han sido mínimos. Se conoce que el espectro de magnitud representa la información del sistema mucho mejor que el espectro de fase comparando desde el punto de vista visual. No obstante, algunos investigadores se han dado a la tarea de explorar este tipo de acercamiento, dando lugar a una caracterización más orientada al trabajo con las señales pero desde el punto de vista espectral. De acuerdo con esto el lenguaje hablado puede ser entonces caracterizado por información ya sea de magnitud o solamente de fase [81].

Uno de los resultados de estas investigaciones fue la obtención de las características modificadas de retraso del grupo (MODGDF) derivadas a partir de la función de igual nombre.

La función de retraso del grupo [81] se define como la derivada negativa de la fase, y puede usarse eficazmente para extraer diversos parámetros del sistema cuando la señal en estudio es una señal de fase mínima. Esto es primordialmente porque el espectro de magnitud de una señal de fase mínima [81], y su función de retraso del grupo se parecen la una a la otra.

Los valores de dicha función que se desvían del valor constante indican el grado de no linealidad de la fase. Para una señal de habla la función de retraso del grupo se calcula usando: [82]

$$\tau_x(\omega) = \frac{|X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)|}{|X(\omega)|^2}$$

Donde los subíndices R e I denotan las partes reales e imaginarias de la transformada de Fourier. $X(\omega)$ y $Y(\omega)$ son las transformadas de $x(n)$ y $nx(n)$, respectivamente. Esta función es de

naturaleza picuda debido a que el tono alcanza niveles máximos, efectos de ruido y de la ventana [82].

Para convertir la función modificada de retraso del grupo en parámetros significativos, la función de retraso del grupo es convertida a características cepstrales utilizando la Transformada Discreta del Coseno (DCT).

$$c(n) = \sum_{k=0}^{N_f} \tau_x(k) \cos(n(2k+1)\pi/N_f)$$

Donde N_f es el orden de la DFT y $\tau_x(k)$ es la función de retraso de grupo. Específicamente se usa la segunda forma de la transformada discreta del coseno (DCT II), aprovechando que tiene propiedades asintóticas, se utiliza la transformación Karhunen -Loeve (KLT) [82].

La DCT se utiliza para eliminar la correlación lineal, lo cual permite el uso de covarianzas diagonales a la hora de modelar el vector de distribución del lenguaje $c(n)$, que es quien brinda las características modificadas de retraso del grupo (MODGDF).

La función modificadora de retraso de grupo ha sido usada en la construcción de reconocedores fonéticos en el procesamiento del habla, [82] así como en la identificación de locutores, [83] y el reconocimiento de sílabas [84].

Para el reconocimiento de lenguaje se emplearon los MODGDF en un sistema LID con un clasificador de máxima verosimilitud (GMM), mostrando que son capaces de identificar el lenguaje tanto a partir de locuciones leídas, [85] como espontáneas [6].

En experimentos posteriores se han usado vectores de dimensionalidad 16 para reconocimiento, necesitando ser reducidos a dos dimensiones para hacer posible la percepción visual de las características. Este trabajo mostró que los rasgos MODGDF mejoran en un 3% el rendimiento de los sistemas LID, con respecto a los rasgos MFCC, y que su combinación (concatenación) es aun mejor. Mostraron además tener mejor funcionamiento en términos de separación de clases cuando se utilizan métricas como Bhattacharya en el proceso de clasificación [86].

Componentes de Modulación de Frecuencia (FM)

Es importante notar que se abre una nueva tendencia a la hora de escoger los rasgos para modelar los lenguajes, puesto que se puede usar para ello tanto la información de magnitud del espectro como información de fase de las señales de habla. Sin embargo, hay que tener cuidado porque los rasgos que se basan en este ultimo tipo de información pueden resultar relativamente inestables, ya que se hace difícil separar exactamente la información de fase.

Este hecho sirvió de motivación para que los investigadores se dieran a la tarea de buscar soluciones. La primera de ellas, recientemente explicada, fue la utilización de la función de retraso de grupo a partir de la que se obtienen los rasgos MODGDF, quienes no presentan este problema, pero su inclusión en los sistemas LID no mostró mejoras estables o significativas en cuanto a funcionamiento [83].

Otra motivación a la búsqueda de nuevas características basadas en la fase fue el análisis del proceso de extracción de los comúnmente usados MFCC, donde se notó que sólo se usa la parte enventanada de magnitud del espectro para producir los coeficientes, dejando simplemente descartada la fase en esa parte del espectro. Es importante notar que esa técnica no es única para el proceso de extracción de los MFCC, sino que también se utiliza en otros métodos de extracción de características basados en magnitudes espectrales. Este es el resultado de que las

teorías clásicas no sugieren que la información de fase tenga una contribución significativa para la inteligencia audible humana [87]. Como respuesta a situación se propone el uso de los componentes de Modulación de Frecuencia (FM) de una señal de habla en un modelo AM-FM [88].

En esta nueva alternativa las resonancias del tracto vocal son representadas como señales AM-FM en un modelo AM-FM. Dicho modelo sugiere la descomposición de señales de voz en una serie de señales modeladas en la frecuencia instantánea y la amplitud. Por eso una señal de habla puede estar representada como la suma de todas las resonancias producidas por el tracto vocal.

Ahora bien, considerando que las distribuciones de frecuencia y tiempo de las señales del habla contienen información acústica que no es visible en la parte lineal del espectro del habla; y que los resultados en ASR [89] han dado indicio de que la parte significativa de la información acústica no puede ser modelada por un modelo acústico que tome como fuente un filtro lineal, se hace evidente entonces, la necesidad de modelar las características no lineales que aparecen en la señal. Estas características, que se basan ya sea el FM o en el AM, proveen información acústica adicional.

Para aislar el componente de frecuencia modulada de cada resonancia de los demás se aplica un filtro de paso de banda a cada componente de resonancia. Para la extracción del componente de frecuencia modulada existen varios métodos, incluyendo al popular algoritmo DESA [90] y a los métodos basados en las transformadas de Hilbert [91].

Sin embargo, las estimaciones de frecuencia modulada producidas por estas técnicas no son suficientemente coherentes debido a picos que salen a relucir ocasionalmente [90, 92]. Consecuentemente la clasificación no puede ser lograda con suficiente exactitud, usando estas estimaciones de frecuencia modulada directamente como las características.

Para producir características basadas en una modulación de frecuencia más consistente fue propuesta una técnica alternativa en [92], en la que la señal de cada banda es modelada por un resonador de segundo orden con todos los polos, del cual se derivan los componentes de frecuencia modulada. Este método evita las variaciones de orden superior de una forma simple y eficaz. Los parámetros del resonador son estimados utilizando predicción lineal y la estimación de frecuencia modulada puede ser derivada desde el ángulo del polo del resonador.

Se presume que un modelo AM-FM sea válido sobre la duración de una ventana centrada en n . El tamaño de la ventana es heurísticamente seleccionado como 25ms con un solapamiento de 15ms. Los componentes de frecuencia modulada de todas las bandas son entonces concatenados para formar un vector de característica de FM [93].

Las características de modulación AM-FM tienen dos ventajas principales comparadas con la representación lineal de los MFCC. Pueden modelar la naturaleza dinámica del lenguaje y dan la apariencia de ser relativamente resistentes al ruido, por lo que logran mejorar el rendimiento y los resultados de los sistemas LID.

“Warping” (arqueo o distorsión) de las características

Los aportes más recientes hechos en esta línea de investigación están orientados a mejorar el rendimiento de los sistemas, ya sea proponiendo nuevas características que aporten algún tipo de información no conocida hasta el momento, o incursionando en los métodos de extracción de las características o en el pre-procesamiento de las señales, en vista de plantear algún cambio que resulte en una disminución del costo computacional o en el aumento del rendimiento de los sistemas. En este aspecto el uso de la técnica “warping” sobre alguna de las características ya

conocidas o incluso sobre la concatenación de varias de ellas, ha sido el mayor aporte en los últimos años.

Este método también conocido en la literatura del procesamiento de imágenes como la igualación del histograma y en la literatura del procesamiento de voz como el mapeo de la distribución acumulativa y fue sugerida como un método para proveer robustez en contra de la incompatibilidad del canal en un sistema de identificación del locutor [94] y los efectos de ruido no lineal en un sistema para reconocimiento robusto del habla [95]. Recientemente se comenzó a usar en los sistemas LID basados en rasgos acústicos a propuesta de Allen, [96] quien demuestra que el “warping” de las características es efectivo no sólo en los ámbitos anteriormente expuestos, sino también como un método de pre mapeo de datos para crear mayor compatibilidad con los clasificadores a usar.

La técnica consiste en trazar un mapa de cada vector de característica para una distribución estándar durante un espacio de tiempo especificado. Para ello se aplica una ventana rectangular corrediza de N muestras, y se calcula el valor nuevo distorsionado para la característica cepstral del centro de la ventana como se muestra en la figura 7, donde la R es el índice del nuevo valor del centro cuando la ventana es ordenada ascendentemente y norminv es la inversa de la función de distribución normal acumulativa.

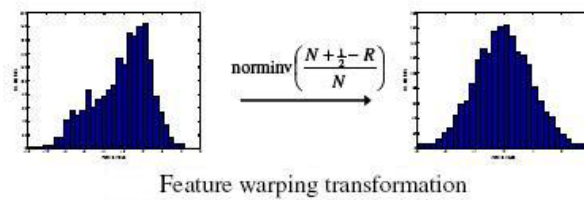


Fig. 7. Proceso de “warping” de características

Los estudios han mostrado que los sistemas de identificación de lenguaje que utilizan esta técnica exhiben un funcionamiento superior en relación a aquellos que utilizan otros métodos incluyendo la Sustracción Media Cepstral (CMS) y la normalización de la media y la varianza.

Se plantea también que el “warping” de características es más conveniente aplicada a los rasgos MFCC o PLP, no proveyendo beneficios cuándo es aplicado a los coeficientes dinámicos (delta y aceleración) o a los SDC.

El “warping” de características puede ser aplicada usando cualquier probabilidad de distribución estándar, pero los mejores resultados se obtienen cuando se utiliza una distribución gaussiana [96].

3.2.2 Clasificadores generativos y discriminativos

Los sistemas de reconocimiento de lenguaje basados en características espectrales operan en función de extraer los rasgos de la señal de voz que mejor caractericen al lenguaje a identificar, y se le aplica algún clasificador que permita delimitar exitosamente cada idioma. En este sentido se han utilizado varias técnicas de clasificación, resultando ser las más exitosas en esta tarea, la modelación por mezclas Gaussianas (GMM), y las máquinas de vectores de soporte (SVM).

GMM

La modelación por mezclas gaussianas es una técnica de clasificación de carácter generativo. Este acercamiento para un clasificador GMM consiste en modelar la distribución de características cepstrales para cada lenguaje para ser reconocido. Un GMM específico al lenguaje estaría dado por:

$$p(x|\lambda) = \sum_{i=1}^M \pi_i \mathcal{N}_i(x)$$

Donde x es un vector de características de dimensión D , $\mathcal{N}_i(x)$ son los componentes de densidades, y π_i son los pesos de la mezcla. Cada componente de densidad es una función Gausiana de D variables con la forma:

$$\mathcal{N}_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \times \exp \left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}$$

Con vector de medias μ_i y matriz de covarianza Σ_i . Los pesos de la mezcla satisfacen la restricción

$$\sum_{i=1}^M \pi_i = 1$$

que asegura que la mezcla sea una función de densidad de probabilidades de aciertos. La parametrización completa de una mezcla de densidad Gausiana específica al lenguaje está dada por los vectores de medias, las matrices de covarianzas, y los pesos de todos los componentes de densidades de la mezcla, lo cual está representado por la notación: $\lambda = \{\pi_i, \mu_i, \Sigma_i\} \quad i = 1, \dots, M$.

En el caso específico del reconocimiento de lenguaje basado en características SDC, los estudios empíricos demostraron [97] que los GMM de altos ordenes ($M = 2048$ y covarianzas diagonales) proveen mejoras sustanciales en el funcionamiento de los sistemas LID.

Los parámetros GMM son estimados por medio del algoritmo de máxima expectación (EM) [98] o por adaptación Bayesiana de un modelo universal (UBM) [99], tal como se hace en el reconocimiento del locutor. La adaptación al modelo universal tiene ventajas computacionales en lo referente a que apresura la marcha de entrenamiento y los cálculos de probabilidad de ambos modelos. Para el reconocimiento de lenguaje, el UBM está típicamente construido entrenando a un GMM a partir de una colección de los datos de entrenamiento de todos los lenguajes disponibles. Recientemente, el entrenamiento discriminativo de parámetros GMM basados en la optimización de la máxima información mutua ha mostrado resultados muy alentadores [100].

Para una tarea general de reconocimiento de lenguaje, un GMM, λ_l , es entrenado para cada uno de los lenguajes deseados, $l = 1, \dots, L$, y se calcula la probabilidad de los vectores de secuencia de características, $X = (x_1, \dots, x_T)$, extraídos de una expresión experimental como:

$$p(X|\lambda_l) = \prod_{t=1}^T p(x_t|\lambda_l)$$

La probabilidad del modelo se calcula asumiendo independencia entre los vectores de características, por lo que el ordenamiento temporal de los vectores de característica no tiene importancia. Aunque, los vectores de característica SDC x_t se extraen durante un intervalo de tiempo multi-tramas y por consiguiente contienen información temporal de la secuencia.

SVM

Las máquinas de vectores de soporte (SVM) son clasificadores flexibles que recientemente han mostrado resultados prometedores en el reconocimiento de lenguaje. Los SVM confían en separar datos en espacios dimensionales altos usando para ello el concepto del máximo margen. [101] Un SVM, $f(Z)$, está constituido por las sumas de una función kernel $K(\cdot, \cdot)$

$$f(Z) = \sum_i \alpha_i K(Z, Z_i) + d.$$

Donde Z es el vector de características de entrada, la sumatoria en i de α_i es $= 0$, y α_i es diferente de 0. Z_i son vectores de soporte que son obtenidos a partir del conjunto de entrenamiento aplicando un proceso de optimización [102]. Los valores ideales del clasificador son 1 o -1, en dependencia de a que clase corresponda el vector de soporte (clase 0 ó clase 1). En el proceso de clasificación, la decisión de la clase se basa en si el valor de $f(Z)$ está por encima o por debajo de un umbral.

El kernel de SVM es un método para comparar vectores de secuencias de características. Dado dichos vectores de dos expresiones, $X = (x_1, \dots, x_{Tx})$ y $Y = (y_1, \dots, y_{Ty})$, el kernel produce una comparación que facilita la discriminación entre los lenguajes de las expresiones. Si las expresiones son del mismo lenguaje, se obtendrá como resultado un valor positivo considerablemente grande; de otra manera, el kernel debería tener como resultado un valor negativo también considerablemente grande.

Para entrenar un sistema de reconocimiento de lenguaje usando SVM se utiliza una estrategia de uno vs todos tal como se muestra la figura 8. El ejemplo que presenta es un sistema para identificar el inglés, por lo tanto, la clase 1 contiene sólo datos referentes al lenguaje objetivo (inglés) y la clase 0 contiene todos los datos del resto de los lenguajes para los que se está entrenando el sistema general (no inglés) puestos en un fondo común. Para el proceso de entrenamiento se utiliza una herramienta de entrenamiento de SVM estándar con un módulo de kernel secuencial. [102] Como resultado se obtiene un modelo que permite reconocer el lenguaje deseado.

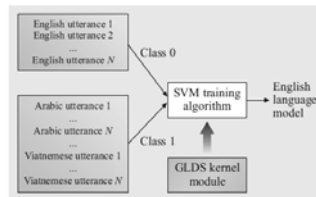


Fig. 8. Estructura general de un entrenamiento SVM basado en la técnica uno vs todos

Un kernel útil para el reconocimiento de lenguaje es el kernel lineal de secuencia discriminante generalizado (GLDS) [103]. Este kernel es simplemente un producto interno entre los promedios de expansiones de alta dimensión de un conjunto de funciones bases $b_j(\cdot)$

$$b(x) = (b_1(x) \cdots b_K(x))^T$$

Donde la K es el número de funciones de base. Típicamente, se usan monomios hasta un grado dado como función base, por ejemplo:

$$b(x) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \\ x_1^2 \\ \vdots \\ x_1^{i_1} x_2^{i_2} \dots x_n^{i_n} \\ \vdots \end{pmatrix}.$$

Quedando la fórmula del kernel: $K_{GLDS}(X, Y) = \bar{b}_x^T \bar{R}^{-1} \bar{b}_y$ y el mapeo de $X \rightarrow bx$ quedaría

$$X \rightarrow \frac{1}{T_x} \sum_{t=1}^{T_x} b(x_t)$$

definido como:

El vector b_y queda definido en manera análoga a la ecuación anterior. R es una matriz de correlación derivada de un conjunto grande de datos provenientes de varios lenguajes y locutores. Usualmente R es una matriz diagonal.

En un sistema general de reconocimiento de lenguaje, se entrenan un conjunto de modelos SVM $\{f_l\}$ para representar cada lenguaje objetivo $l = 1, \dots, L$. Dada una expresión de prueba, se extrae la secuencia de vectores X , y se mapean en la expansión promediada, convirtiéndose en un conjunto de puntuaciones $s_l = f_l(bx)$. Las puntuaciones SVM s_l son transformadas a s'_l prima usando un cálculo de tipo de proporción de probabilidad:

$$s'_l = s_l - \log \left(\frac{1}{M-1} \sum_{j \neq l} e^{-s_j} \right)$$

La transformación anterior supone que las puntuaciones SVM se comportan como las proporciones de probabilidad logarítmicas [9, 104].

En la actualidad existen varias formas simples de implementar un sistema SVM para el reconocimiento de lenguaje. Por ejemplo, las puntuaciones del reconocimiento de lenguaje pueden ser reducidas a un solo producto interno. También, el entrenamiento del modelo de lenguaje puede ser simplificado usando un kernel lineal.

3.2.3 Fusiones

En los últimos años las investigaciones han mostrado que existe una tendencia general a hacer fusiones dentro de los sistemas LID. Estas fusiones se hacen en diferentes niveles dentro del sistema. Entre los casos más frecuentes se encuentran la concatenación de varios tipos de características o rasgos y la fusión de clasificadores. Tanto en un caso como en el otro, el principal objetivo que se persigue es mejorar el rendimiento de estos sistemas. A continuación se presentan algunos casos:

Fusionando rasgos

La figura 9 muestra un ejemplo claro del uso de fusión de características para un sistema de identificación de lenguajes. Este vector de rasgos que combina tanto información de magnitud como de fase fue propuesto en el 2006 por Allen [104] en la búsqueda de una mejor representación de los lenguajes.

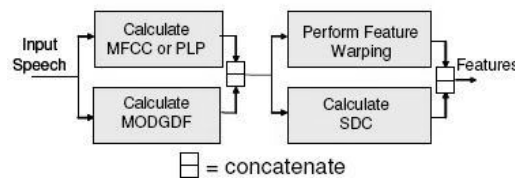


Fig. 9. Representaciones de la fusión de características

El proceso de extracción de características fue el siguiente. Se extrajeron los coeficientes basados en magnitudes (PLP o MFCC) y se calcularon los coeficientes MODGDF combinándose ambas informaciones por medio de la concatenación de las características. Una vez creado el vector se procedió a aplicar la técnica “warping” sobre el y de forma paralela se calcularon los SDC usando un nuevo método basado en regresión. Finalmente los SDC son concatenados a las características obteniéndose el vector de rasgos final. Como resultado de este trabajo se obtuvo un 80.9% de aciertos para la combinación con los MFCC y un 76.4% combinando con rasgos PLP.

Fusionando métodos

Para este análisis es conveniente dividir la discusión en dos tipos de fusión, fusión basada en las reglas y fusión basada en los clasificadores. La fusión basada en las reglas se basa en aplicar una fórmula fija a los resultados, como una regla de suma ponderada o la regla del producto norma, donde los pesos son uniformes o se eligen empíricamente a partir de un conjunto de los datos de desarrollo. Un ejemplo de una fusión basada en las reglas para combinar las puntuaciones resultantes en un sistema PPRLM. Por su parte la fusión basada en clasificadores utiliza un conjunto de datos para entrenar un clasificador a partir de un vector de características construido con las puntuaciones individuales del sistema. En teoría, el conjunto de estadísticas entre los resultados (los elementos del vector de características) puede ser aprovechado por el clasificador para producir el desempeño total, que es superior al alcanzado con una regla como respaldo del sistema.

Un método para la fusión de las puntuaciones de un sistema LID PPRLM es utilizar la regla de producto propuesta por Zissman [23]. Este asume un reconocedor de idioma PPRLM que

utiliza K reconocedores fonéticos para detectar L idiomas. Para cada entrada de prueba, se genera un conjunto de KL (lineal) puntuaciones, con los que cada uno de los reconocedores fonéticos produce L resultados. Un posible inconveniente de usar la regla de producto para fusión es que una sola puntuación cercana a cero puede tener un impacto indeseable en el resultado final. A pesar de esto, este método ha demostrado ser una forma eficaz de combinar los resultados PPRLM.

La fusión basada en Clasificadores en los sistemas LID PPRLM fue propuesta inicialmente por Yan y Bernard [105] usando redes neuronales y por Zissman [106] usando clasificadores Gausianos, reportando este último que el rendimiento obtenido en este caso, fue superior al obtenido con el enfoque de fusión por regla de producto. La figura 10 muestra un diagrama de bloques de la fusión basada en clasificadores Gausianos para el reconocimiento de idioma. El método consiste en la normalización de un análisis discriminante lineal (LDA) seguido por una clasificación Gausiana. Aunque esta arquitectura no garantiza la obtención de los resultados óptimos para todas las aplicaciones de reconocimiento de idiomas, experimentos hechos han demostrado, con fiabilidad, que el rendimiento de este acercamiento es superior. Otros investigadores han reportado resultados usando redes neuronales como clasificador de respaldo para la fusión [107].

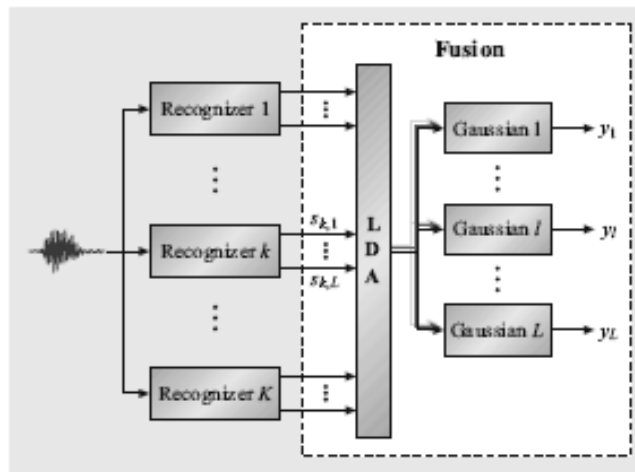


Fig. 10. Fusión basada en un clasificador gaussiano para reconocimiento del lenguaje

El diagrama de bloques genérico de la figura 10 muestra un conjunto reconocedores de lenguaje básico (por ejemplo, GMM, SVM, PRLM, PPRLM), cada uno de ellos produce una clasificación o un vector de puntuaciones representado una hipótesis de lenguaje para cada expresión de entrada dado. Estos resultados se utilizan para formar los elementos del vector característico que se transforma a través de LDA, método en el que el vector característico se proyecta en un espacio dimensional reducido con el objetivo de aprovechar al máximo la separabilidad entre las clases y reducir al mínimo la variabilidad dentro de las clases. En la figura, el espacio transformado es de dimensión $L-1$, donde L es el número de clases (lenguajes) representados en los datos de entrenamiento de fondo. El vector transformado también posee la propiedad de que sus valores no están correlacionados. Es importante tener en cuenta que los diferentes reconocedores pueden producir vectores de puntuaciones de diferentes tamaños y que

los resultados pueden representar las probabilidades de cualquiera de los lenguajes, no sólo de los lenguajes objetivos.

Sistemas LID sobre un marco Jerárquico

Ahora bien, las técnicas de fusión existentes, algunas veces, experimentan dificultad para mejorar el rendimiento cuando el número de lenguajes y características aumenta [108], porque todas las hipótesis de lenguaje son examinadas en un solo nivel. Esto quiere decir que la variación de las distancias entre los lenguajes en espacios diferentes de características no es suficientemente considerada.

Investigaciones recientes han propuesto un sistema LID montado sobre una plataforma jerárquica (HLID) [108]. Este enfoque de clasificación de varios niveles asegura que el hiperplano más discriminativo en el espacio de característica sea usado en cada nivel de clasificación, y por eso logra el mejor rendimiento. Para construir la estructura jerárquica óptima se desarrolla una técnica no supervisada, que agrupe los lenguajes según las separaciones entre grupos en términos de rasgos disponibles o sistemas LID primarios.

En esta plataforma las hipótesis de lenguaje son agrupadas jerárquicamente para formar una estructura de árbol como se muestra en la figura 11. En el árbol, cada hoja es una hipótesis individual de lenguaje, y otros nodos son grupos de lenguajes conteniendo las hipótesis de lenguaje de sus nodos hijos. La expresión de prueba de lenguaje es clasificada nivel por nivel, según las características más discriminativas en cada nivel [108]. Este plan de selección de característica fue reemplazado por un plan de fusión más adelantado en [109], donde varios sistemas LID primarios son fusionados para obtener una decisión de clasificación más precisa en cada nivel. Las flechas gruesas en la figura 11 exhiben un ejemplo de un camino de clasificación.

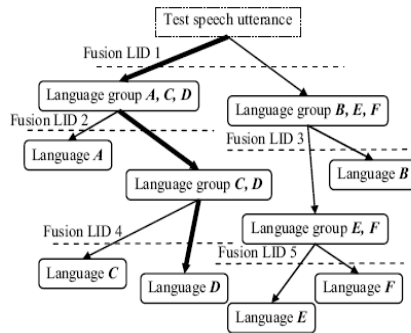


Fig. 11 Proceso jerárquico de identificación del idioma

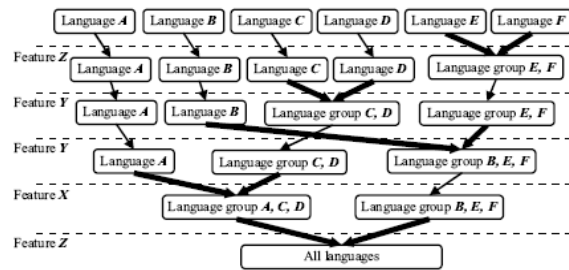


Fig. 12 Proceso de agrupamiento en un LID Jerárquico

Para crear la estructura jerárquica de clasificación, y seleccionar la característica más discriminativa en cada nivel se propuso un proceso de agrupamiento aglomerativo sin supervisión con el objetivo de obtener el hiperplano más discriminativo en el espacio de características para cada nivel de clasificación [108]. La figura 12 muestra como los lenguajes más similares o grupos de lenguajes se combinan para formar mayores grupos, nivel por nivel, basándose en las medidas de distancias seleccionadas. Este proceso de agrupamiento asegura que en cada nivel las distancias entre grupos de lenguaje son siempre mayores que las distancias entre los lenguajes dentro de un grupo dado de lenguaje. Además los modelos del grupo de

lenguaje son más precisos y robustos que los modelos individuales de lenguaje, puesto que los lenguajes dentro de un grupo comparten características similares. En otras palabras, están próximos el uno al otro en el espacio de características. Por consiguiente, un mayor número de datos para entrenar (o sea de todos los lenguajes dentro del grupo) ayuda a aumentar la precisión en el modelo de entrenamiento.

Inicialmente, fue usada una medida basada en funciones centradas en la lejanía entre un par de lenguajes o grupos [108]. Considerando el alto costo computacional y el requisito de un número abrumador de datos de desarrollo, esta medida basada en funciones de distancia fue reemplazada por medias de distancias como Cross Likelihood Ratio (CLR) o divergencia Kullback-Leibler [109]. Con ambas medidas se logra un mayor rendimiento en la clasificación, con menor costo computacional.

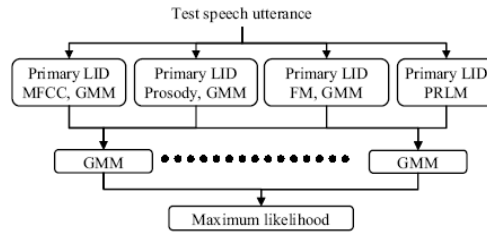


Fig. 13 Fusión de sistemas primarios en uno de los niveles de clasificación

La principal ventaja de este enfoque radica en las posibilidades que brinda a la hora de fusionar los resultados. Es decir, que se pudiera perfectamente identificar el tipo de rasgos y de clasificador primario que mejor rendimiento tiene para cada lenguaje y a partir de este construir un gran sistema LID basándose en esta plataforma. Durante el proceso de prueba la señal en cuestión será pre-procesada y sus rasgos sometidos a comparación con cada uno de los modelos de lenguajes. Como resultado de la clasificación de cada nivel se obtendrá la hipótesis de lenguaje o grupo de lenguajes que mayor valor de verosimilitud tenga asociado (ver figura 13) [93].

3.3 Caracterización en espacio-vectorial

La modelación del espacio vectorial “VSM” se ha convertido en una herramienta estándar en los sistemas de Recuperación de Información desde su introducción hace varias décadas. Este acercamiento usa un vector para representar un documento de texto o una búsqueda. También ha sido aplicado a la Categorización de Texto en la que los vectores de entrenamiento son usados para diseñar una colección de clasificadores de temáticas. Este acercamiento ha sido recientemente propuesto para la identificación de lenguajes LID.

Suponiendo que una expresión X está representada por una secuencia de vectores de características del habla O , esta es tokenizada en un documento hablado, $d(X)$, consistente en una serie de I unidades acústicas, $d(X) = \{t_1, \dots, t_i, \dots\}$, donde cada unidad es dibujada desde un inventario universal $U = \{u_1, \dots, u_j, \dots\}$ de J letras acústicas tomadas en común de todos

los lenguajes hablados que fueron considerados, tales que $t_i \in U$. Entonces se puede establecer una colección de palabras acústicas AWs agrupando unidades que ocurren consecutivamente para obtener un vocabulario de M distintas palabras, $W = \{w_1, \dots, w_m, \dots, w_M\}$, tales que cada w_m puede ser una palabra de una sola letra como (u_i), una palabra de dos letras como (u_juk), de tres letras como (u_jukul), y así sucesivamente. Generalmente el tamaño del vocabulario, M, es igual al número total de los n-gramas patrones necesarios para formar las palabras, es decir, $M = J + J \times J + J \times J \times J$ si la longitud de la palabra acústica está por encima de los 3 tokens es considerado válido.

Posteriormente se puede usar alguna forma de función $f(w_m)$ como LSI, para evaluar cuán significativa es tener la palabra w_m en el documento $d(X)$. A partir de ese momento se puede establecer un vector de características de M dimensiones, $v = [f(w_1), \dots, f(w_m), \dots, f(w_M)]$ en el que x^T denota la traspuesta del vector x para cada documento hablado.

Es de notar que se necesita un número de unidades fundamentales suficientes para cubrir la variación acústica en el espacio del sonido. Sin embargo una J grande resultaría en un vector de características con una muy alta dimensión si se quisiera cubrir como la combinación de varias unidades cuando formamos palabras acústicas. Por ejemplo, con un valor moderado de $J = 256$, se tiene $M = 65792$ incluso cuando solo se consideren palabras mayores o iguales a dos letras. Esto es ya una alta dimensionalidad no comúnmente utilizada por los algoritmos de procesamiento de locutores y lenguajes. Finalmente, Un clasificador basado en vectores evalúa una precisión del ajuste, o una función score $Sl(v) = S(v|\lambda_l)$ entre un vector v dado y el modelo del l-ésimo lenguaje hablado λ_l para tomar una decisión. Cualquier clasificador basado en vectores puede ser usado para diseñar sistemas de identificación y verificación de locutores. Una buena aproximación sería usar una función discriminante lineal como producto interno, para evaluar las puntuaciones basándose en los vectores:

$$S(v|\lambda_l) = \alpha \gamma_l^T \cdot v$$

donde γ_l es el vector de peso dependiente del idioma de igual dimensión que v , con cada atributo representando la contribución de su probabilidad individual de n-gramas para los scores globales del idioma. El vector del documento hablado v en la ecuación anterior tiene una alta dimensionalidad cuando se incluyen hasta patrones triples. Cuando son usados varios inventarios de modelos, como en P-PRLM, para producir una colección de F vectores documentos $\{v_f, f = 1, \dots, F\}$, si estos F vectores son unidos y se puede establecer un gran vector documento, o supervector $v = \{v_1^T, \dots, v_F^T\}^T$. Se ha mostrado que tales supervectores tienen más poder de discriminación que los vectores singulares del documento, para el reconocimiento de lenguaje.

El término de pesos es ampliamente usado para dar valor a los atributos en un vector documento teniendo en cuenta la frecuencia de ocurrencia de cada atributo. Es interesante reparar en que los patrones de atributos que ocurren a menudo en algunos documentos, pero no tanto en otros, le da un gran poder de indexación a estos documentos. Por otra parte, los patrones que ocurren muy a menudo en todos los documentos poseen muy poco poder de indexación. Esta propiedad, conduce a un número de términos de peso planificados, tales como *tf-idf* ("term-frequency-inverse document frequency") [110] y LSI [111] los cuales son comunes para la recuperación de información, [112] la determinación del recorrido de una llamada de

lenguaje natural [113] y la categorización de texto. [114] VSC es motivado por las mismas ideas, con la intención de derivar los pesos que discriminan entre los lenguajes usando vectores de características resultantes con altas dimensiones.

Después de representar el documento hablado como un vector de las estadísticas de las palabras acústicas, la identificación del lenguaje se convierte en un problema de clasificación basado en vectores. En la literatura hay varios diseños de clasificadores de aprendizaje automático para la clasificación de vectores de altas dimensiones, usándose las técnicas convencionales tales como SVM, arboles de clasificación y regresión CART [115] y redes neuronales artificiales ANN [116] para entrenar los clasificadores basados en vectores. La respuesta final puede ser resuelta de la forma

$$\hat{l} = \underset{l}{\operatorname{argmax}} S(v|\lambda_l)$$

También se puede considerar un mecanismo indirecto de vectorización basado el cálculo de las puntuaciones para todas las clases de interés, obtenidas de una combinación de clasificadores. Estas puntuaciones caracterizan la distribución de las salidas de la variedad de clasificadores, proveen el poder discriminatorio significativo entre las clases, y pueden ser agrupadas juntos para formar supervectores de puntuaciones describiendo el comportamiento global de una distribución de puntuaciones sobre diferentes clasificadores para todas las clases competentes.

Un ejemplo de tal supervector es concatenar las puntuaciones de los estados de HMM de todos los modelos competentes para formar un vector de puntuaciones general. Este método ha sido utilizado para tener un buen poder discriminatorio en el reconocimiento esporádico de letras. Dicho supervector de puntuaciones es obtenido generalmente de un conjunto infinito de clasificadores combinados, su dimensionalidad es mucho mayor que la de los vectores de los documentos hablados discutidos anteriormente [10].

Sistemas LID usando VSM

Los sistemas LID que se han desarrollado bajo esta tendencia están montados sobre una plataforma de dos partes. Una que depende fundamentalmente del pre procesamiento que se les dé a las muestras y otra que es la referida al proceso de clasificación VSM visto anteriormente.

Una primera aproximación a estos sistemas sería la caracterización de un lenguaje hablado usando distribuciones de probabilidad de características espectrales en forma de unidades lingüísticamente definidas como fonemas y silabas. Aquí los modelos fonéticos se usan para convertir expresiones de lenguaje en secuencia de símbolos fonéticos a partir de las puntuaciones de probabilidad acústicas resultantes. Se crea entonces un modelo de lenguaje interpolado n-grama basado en fonemas para cada lenguaje, y se derivan las puntuaciones fonotácticas (más o menos lo mismo que PRLM).

Llegado este punto se intenta derivar una colección universal de fonemas para cubrir todos los sonidos de los lenguajes descritos en un inventario fonético internacional. Es un empeño desafiante porque se necesitó tomar muestras de una gran colección de lenguajes, teniendo en cuenta que los sonidos se traslapan considerablemente a través de los lenguajes. Una aproximación posible es usar un conjunto de fonemas de varios lenguajes para formar un super inventario fijo, denominado inventario aumentado de fonemas (API). Anticipamos que un inventario fonéticamente bueno debe cubrir tantos lenguajes específicos como sea posible. Este

método puede ser efectivo cuando los fonemas de todos los lenguajes dirigidos a sectores específicos forman un conjunto cerrado [16].

Al aplicarle este concepto de unidad universal a la discriminación de lenguaje, este se puede relacionar con el hecho de que la entropía del idioma puede estar eficazmente reducida cuando se calculan las estadísticas de alto nivel de las letras. Si se puede tokenizar el lenguaje con un conjunto dócil de letras habladas y se desarrollan modelos para descifrarlos a partir de expresiones habladas, en ese entonces es claro que las estadísticas de estas letras habladas y su co-ocurrencias pueden usarse para discriminar un lenguaje hablado de otro.

Una generalización interesante por medio de unidades acústicas es representar cualquier expresión hablada con un vector de característica de alta dimensión, donde cada elemento de sonido acarrea estadísticas de co-ocurrencia.

Haizhou Li en su trabajo hace la suposición de que las características de sonido de todos los lenguajes hablados puedan estar cubiertas por un conjunto universal de unidades acústicas derivadas automáticamente sin vínculo directo con las definiciones fonéticas. Estos correspondientes modelos (HMM), llamados modelos de segmentos acústicos (ASMs), pueden ser usados para descifrar expresiones habladas en forma una serie de tales unidades [117]. Estos modelos se han usado para hacer un glosario acústico para el reconocimiento esporádico de palabras con un alto nivel de exactitud. En el glosario acústico, los ASMs forman letras habladas, y sirven como una base de la cual se derivan los vectores de características para la constitución de los documentos hablados y los clasificadores para la identificación automática de lenguaje. La identificación de lenguajes es entonces tratada como un problema de categorización de texto en el que la extracción de características y el diseño de los clasificadores son dos de los principales componentes de investigación [118].

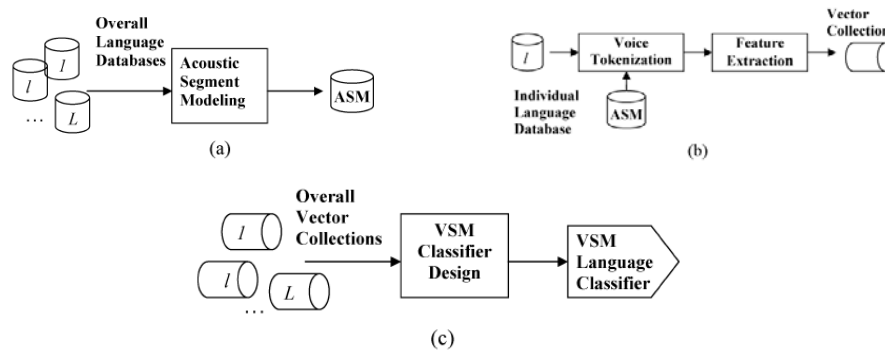


Fig 14. Diagrama en bloques de las diferentes etapas de un sistema LID ASM-VSM

La figura 14 muestra de forma muy general un diagrama de los diferentes procesos que se llevan a cabo en un sistema LID que usa ASM-VSM. La primera etapa (a) consiste en la confección del API o en el caso del ejemplo, el ASM. La segunda fase (b) consiste en la transcripción de las señales de entrenamiento al documento hablado y a partir de este el proceso de extracción de características, hasta obtener el conjunto de vectores con el que se va a entrenar el clasificador discriminativo para crear el modelo del lenguaje durante la última fase (c).

Otra aproximación de este enfoque fue la presentación de un sistema LID pero sin la generalización universal, es decir que se hace la transcripción fonética usando las bases de datos

dependientes de cada idioma, lo cual se conoce como PPR-VSM, puesto que en su primera fase trabaja cual si fuera un reconocedor fonético de un sistema PRLM de los vistos en el epígrafe 3.1.1. Bin Ma en su artículo usó las técnicas de descomposición en valores singulares SVD y análisis de componentes principales PCA para reducir la dimensionalidad de los vectores y luego aplicó un clasificador generativo de GMM para crear los modelos de los lenguajes y estimar los valores de verosimilitud en la fase de prueba [119].

4 Experimentación

Para tener una panorámica más completa del estado del arte de este tema, se incluyeron los resultados experimentales obtenidos a partir de un sistema LID primario para la identificación del idioma Inglés. La implementación fue hecha en Matlab R2006b siguiendo el seudocódigo propuesto por Campbell y sus colegas en [9].

Tanto los datos de entrenamiento como los de prueba fueron tomados de la base de datos CSLU que contiene expresiones telefónicas en 21 idiomas. Durante el entrenamiento fueron procesadas 229 expresiones del idioma Inglés para construir el modelo target y 525 expresiones para el modelo de los non-target. Estas últimas eran señales de idioma Portugués y Alemán. Los datos de prueba fueron grupos de 100 señales de 30, 10 y 3 seg., compuestos de 5 idiomas: alemán, ruso, inglés, español y portugués (20 por cada idioma).

5 Descripción y resultados

Para el sistema SVM, se extraen los rasgos SDC siguiendo la representación de 7-1-3-7 para los parámetros N-d-p-k respectivamente [9]. Después de extraídos los rasgos, se eliminan las tramas de silencio y cada rasgo es normalizado con media 1 y varianza 0 sobre la base de cada expresión. Como resultado de este proceso se obtuvo una secuencia de vectores de características de dimensión 49 para cada expresión.

En un segundo paso estos vectores de rasgos son expandidos usando el kernel GLDS [9]. Como función de expansión se usaron todos los posibles monomios de grado 3 obteniendo como resultado vectores con dimensión 18,523. Para la optimización del SVM se usó la herramienta SVMTool [101]. Los resultados obtenidos se muestran en la siguiente tabla.

	<i>Pos</i>	<i>FP</i>	<i>Neg</i>	<i>FN</i>
3 seg	11	20	60	9
10 seg	7	21	59	13
30 seg	12	8	72	8

Los resultados obtenidos en esta primera fase experimental indican que, como ya se había comprobado en investigaciones anteriores, los rasgos SDC aportan información de primera clase para la identificación de los idiomas [2, 3].

Se conoce que los resultados están sujetos también a determinadas condiciones experimentales que pudieron influir en la obtención de los mismos. Entre estas podemos

destacar: La insuficiente cantidad de datos tanto para entrenamiento como para prueba, pues estos sistemas necesitan de grandes volúmenes de datos (en el orden de horas) para lograr modelos de target efectivos. Y que el orden de la dimensión que se logra con la función de expansión implementada dentro del kernel GLDS fue de 18 523, no así el que teníamos previsto 12 000.

Fuera de esto, se considera que haciendo los ajustes pertinentes, esta implementación puede servir para establecer como sistema de línea base, para las futuras comparaciones.

6 Conclusiones

En la actualidad existen tres corrientes fundamentales dentro del área del reconocimiento del lenguaje. Una primera aproximación orientada al reconocimiento fonético. Los sistemas LID desarrollados bajo este enfoque se centran en la identificación de los fonemas y a partir de ellos crear los modelos de los idiomas a identificar. Los principales exponentes de esta tendencia son el PRLM, el P-PRLM y el PPR, cada uno con sus características propias. Estos sistemas tienen como desventaja fundamental que dependen de un alto número de señales para el entrenamiento y que cada una de ellas debe ser escuchada y etiquetada por un especialista lingüista, por lo cual resulta muy costosa su implementación y aplicación.

Un segundo acercamiento lo constituyen los sistemas que no utilizan el reconocimiento fonético, sino que se basan en la extracción de características representativas de los lenguajes a partir de las cuales crea los modelos con la ayuda de un clasificador. Los rasgos extraídos pueden estar en función de representar las características de magnitud de espectro, la información de fase del mismo o incluso una mezcla de ambos como es el caso de las fusiones. Los clasificadores usados pueden ser generativos (GMM) o discriminativos (SVM), obteniéndose con ambos buenos resultados. Es importante destacar que aunque el costo computacional de estos sistemas es mucho menor, y no tiene dependencia de especialista alguno, aun no se han logrado superar los resultados con respecto a los sistemas con reconocimiento fonético.

El tercer y más reciente acercamiento es como una fusión entre los dos acercamientos. Parte de hacer un reconocimiento fonético y a partir de este punto utiliza algunas técnicas de recuperación de información y categorización de texto para hacer una transcripción fonética y convertir el documento hablado en un conjunto de supervectores para cada lenguaje. A partir de este punto, los sistemas se comportan como un proceso de clasificación directa sobre los rasgos. Un clasificador es entrenado con los datos de cada idioma y se crean los modelos pertinentes. Este enfoque aunque ha mostrado muy buenos resultados, también tiene la desventaja de depender de bases de datos para hacer esta transcripción fonética, proceso que encarece notablemente su aplicación y puesta en práctica.

Como resultado final de la investigación se estima que en el área de los sistemas con reconocimiento fonético no hay mucho más que hacer. Los avances en esta área están en función de abaratar los costos de entrenamiento, tratando de minimizar el proceso de etiquetado de las señales. A pesar de esto, estos sistemas se mantienen con los mejores resultados en cuanto a eficacia a la hora de identificar un lenguaje. Por otra parte existe una tendencia generalizada a hacer transformaciones orientadas fundamentalmente a la obtención de rasgos que sean más

robustos y discriminativos, ya sea sobre la base de fusionar los existentes, o incursionar en los procesos de extracción. La inclusión recientemente de la información de fase de la señal de habla dentro de los vectores de características propiciando resultados alentadores, da una muestra clara de que aún quedan cosas por hacer en esta área de investigación.

Referencias bibliográficas

1. Reyes, A.L.: "Un Método para la Identificación Automática del Lenguaje Hablado Basado en Características Suprasegmentales". México (2007)
2. Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller, J.R.: "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features". International Conference on Spoken Language Processing. CO, ISCA, Denver (2002) pp. 33-36, 82-92
3. Singer, E., P.A. Torres-Carrasquillo, Gleason, T.P., Campbell, W.M., Reynolds, D.A.: "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition". Proc. Eurospeech, Geneva, Switzerland, ISCA (1-4 September 2003) pp. 1345-1348
4. Alcaraz-Varó, E., Martínez-Linares, M.A.: *Diccionario de Lingüística Moderna*. In: Ariel, S.A.C. (ed.). ISBN: 84-344-0510-5, Barcelona, España (1997)
5. Abercrombie, D.: *Elements of General Phonetics*. Edinburgh University Press (1967)
6. Muthusamy, Y.K., Cole, R.A., Oshika, B.T.: The OGI multilanguage telephone speech corpus. ICSLP (1992) pp. 895-898
7. Torres-Carrasquillo, P.A., Reynolds, D.A., Deller, J.R.: Language identification using Gaussian mixture model tokenization. ICASSP (2002) pp. 757-760
8. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker recognition. IEEE Signal Processing Letters **13** (2006) 308-311
9. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-Carrasquillo, P.A.: Support Vector Machines for Speaker and Language Recognition Computer Speech & Language **20** (2004) pp 210-229
10. Benesty, J., Sondhi, M.M., Huang, Y.: "Springer Handbook of Speech Processing". Springer-Verlag, Berlin Heidelberg (2008)
11. Cimarusti, D., Ives, R.B.: Development of an automatic identification system of spoken languages: Phase 1. ICASSP (1982) pp. 1661-1663
12. Zissman, M.A.: Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models ICASSP (1993) pp. 399-402
13. Wong, K.K., Siu, M.: Automatic language identification using discrete hidden Markov model. Interspeech (2004) pp. 1633-1636
14. Muthusamy, Y.K.: *A Segmental Approach to Automatic Language Identification* . Oregon Graduate Institute of Science and Technology, Beaverton (1993)
15. Campbell, W.M., Singer, E., Torres-Carrasquillo, P.A., Reynolds, D.A.: Language recognition with support vector machines. Odyssey: The Speaker and Language Recognition Workshop (2004) pp. 41-44
16. Hazen, T.J.: Automatic Language Identification Using a Segment-Based Approach. MIT, Cambridge (1993)
17. Crystal, D.: *A Dictionary of Linguistics and Phonetics*. Basil Blackwell, Oxford: (1985)
18. Ladefoged, P.: Vowels and Consonants: An Introduction to the Sounds of Languages. Blackwell, Oxford (2005)

19. Reproduction of the International Phonetic Alphabet, <http://www.langsci.ucl.ac.uk/ipa/index.html> (2005)
20. Maddieson, I.: *Sound Patterns of Language*. Cambridge Univ. Press, Cambridge (1984)
21. Traill, A.: Phonetic and Phonological Studies of of !Xó'o bushman. *Quellen zur Khoisan-Forschung* 1, Hamburg (1985)
22. House, A.S., Neuberg, E.P.: Toward automatic identification of the languages of an utterance: preliminary methodological considerations. *J. Acoust. Soc. Am* (1977) pp 708-713
23. Zissman, M.A.: Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. Speech Audio Process.* (1996) pp 31-44
24. Schultz, T., Waibel, A.: Language independent and language adaptive acoustic modeling for speech recognition. *Speech Commun* (2001) pp (1-2), 31-51
25. Black, A.W., Schultz, T.: Speaker clustering for multilingual synthesis. *ISCA Tutorial and Research Workshop on Multilingual Speech and Language* (2006)
26. Hazen, T.J., Zue, V.: Segment-based automatic language identification *Acoust. Soc. Am.* (1997) pp 2323-2331
27. Adda-Decker, M., Antoine, F., Mareuil, P.B.d., Vasilescu, I., Lamel, L., Vaissiere, J., Geoffrois, E., Liénard, J.-S.: Phonetic knowledge, phonotactics and perceptual validation for automatic language identification. *International Congress of Phonetic Sciences* (2003)
28. Mareuil, P.B.d., Corredor-Ardoy, C., Adda-Decker, M.: Multi-lingual automatic phoneme clustering. *Int. Congress Phonetic Sci* (1999) pp. 1209-1213
29. Corredor-Ardoy, C., Gauvain, J.L., Adda-Decker, M., Lamel, L.: Language Identification with Language-independent Acoustic Models.: *Eurospeech* (1997) pp. 355-358
30. Ma, B., Li, H.: Spoken language identification using bag-of-sounds. *International Conference on Chinese Computing* (2005)
31. Kirchhoff, K., Parandekar, S.: Multi-stream statistical language modeling with application to automatic language identification. *7th European Conference on Speech Communication and Technology Proceedings of Eurospeech* (2001) pp. 803-806
32. Blevins, J.: The syllable in phonological theory. *The Handbook of Phonological Theory*, Vol. 1. Blackwell Handbooks in Linguistics, Oxford (1995)
33. Kenstowicz, M.: *Phonology in Generative Grammar*, Vol. 7. Blackwell Textbooks in Linguistics, Oxford (1994)
34. Zhu, D., Adda-Decker, M., Antoine, F.: Different size multilingual phone inventories and context-dependent acoustic models for language identification,. *Interspeech* (2005) pp. 2833-2836
35. Berkling, K., Zissman, M., Vonwiller, J., Cleirigh, C.: Improving accent identification through knowledge of English syllable structure. *5th International Conference on Spoken Language Processing* (1998) pp. 89-92
36. Grabe, E., Low, E.L.: Durational variability in speech and the rhythm class hypothesis. *Laboratory Phonology*, Mouton de Gruyter, Berlin (2002)
37. Dauer, R.M.: Stress-timing and syllable-timing reanalysed. *J. Phonet* (1983)
38. Rouas, J., Farinas, J., Pellegrino, F., Andre-Obrecht, R.: Modeling prosody for language identification on read and spontaneous speech. *ICASSP*, Vol. 6 (2003) pp. 40-43,
39. Rouas, J., Farinas, J., Pellegrino, F., André-Obrecht, R.: Rhythmic unit extraction and modelling for automatic language identification. *Speech Commun* (2005) pp 436-456
40. Tong, R., Ma, B., Zhu, D., Li, H., Chang, E.S.: Integrating acoustic, prosodic and phonotactic features for spoken language identification. *ICASSP* (2006) pp. 205-208
41. Schultz, T., Rogina, I., Waibel, A.: Experiments with LVCSR based language identification. *Speech Symposium SRS XV* (1995)

42. Schultz, T., Rogina, I., Waibel, A.: LVCSR-based language identification. ICASSP (1996) pp. 781- 784
43. Matrouf, D., Adda-Decker, M., Lamel, L., Gauvain, J.: Language identification incorporating lexical information. 5th International Conference on Spoken Language Processing (1998) pp. 181-184
44. Hieronymus, J., Kadambe, S.: Robust spoken language identification using large vocabulary speech recognition. ICASSP (1997) pp. 779-782
45. Chelba, C.: *Exploiting Syntactic Structure for Natural Language Modeling* Vol. Ph.D. thesis. Johns Hopkins University, Baltimore (2000)
46. Wang, W., Harper, M.: The SuperARV language model: Investigating the effectiveness of tightly integrating multiple knowledge sources. Conference of Empirical Methods in Natural Language Processing (2002) pp. 238-247
47. Lee, C.-H., Huo, Q.: On adaptive decision rules and decision parameter adaptation for automatic speech recognition. Proc. IEEE **88** (2000) 1241-1269
48. Lee, C.-H., Soong, F.K., Paliwal, K.K. (eds.): *Automatic Speech and Speaker Recognition: Advanced Topics* Kluwer Academic, Dordrecht (1996)
49. Gauvain, J.L., Lamel, L.: Large-vocabulary continuous speech recognition: advances and applications. Proc. IEEE **88** (2000) 1181-1200
50. Zissman, M., Singer, E.: Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'94), Vol. 1, Adelaide, Australia (1994) pp. 305-308
51. Hazen, T.J., Zue, V.W.: Automatic language identification using a segment-based approach. *Eurospeech '93*, Vol. 2 (1993) pp. 1303-1306
52. Tucker, R.C.F., Carey, M.J., Paris, E.S.: Automatic language identification using sub-words models. *ZCASSP '94*, Vol. 1 (1994) pp. 301-304
53. Nakagawa, S., Seino, T., Ueda, Y.: Spoken language identification by ergodic HMMs and its state sequences. *Electron. Commun. Japan, Pt. 3* **77** (1994) 70-79
54. Nakagawa, S., Ueda, Y., Seino, T.: Speaker-independent, textindependent language identification by HMM? : *ICSLP '92*, Vol. 2 (1992) pp. 1011-1014
55. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models,. IEEE Trans. Speech Audio Process. **3** (1995) 72-83
56. Vapnik, V.: *The Nature of Statistical Learning Theory* Springer, Berlin, Heidelberg (1995)
57. Campbell, W.M., Gleason, T., Navratil, J., Reynolds, D., Shen, W., Singer, E., Torres-Carrasquillo, P.A.: Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recongnition evaluation. IEEE Odyssey Speaker and Language Recognition Workshop (2006)
58. Hazen, T.J., Zue, V.W.: Recent improvements in an approach to segment-based automatic language identification. *ICASSP '94*, Vol. 4 (1994) pp. 1883-1886
59. Yan, Y., Barnard, E.: An approach to automatic language identification based on language-dependent phone recognition. *ICASSP '95*, Vol. 5 (1995) pp. 3511-3514
60. Lamel, L.F., Gauvain, J.-L.: Identifying non-linguistic speech features. *Eurospeech '93*, Vol. 1 (1993) pp. 23-30
61. Muthusamy, Y., Berkling, K., Arai, T., Cole, R., Barnard, E.: A comparison of approaches to automatic language identification using telephone speech. *Eurospeech '93* **2** (1993) pp. 1307-1310.
62. Yin, B., Ambikairajah, E., Chen, F.: "Combining Cepstral and Prosodic Features in Language Identification". IEEE (2006)

63. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions* **28** (1980) pp 357 - 366
64. Mammone, R.J., Zhang, X., Ramachandran, R.P.: "Robust speaker recognition: a feature-based approach,". *Signal Processing Magazine, IEEE* **13** (1996) pp. 71 - 129
65. Pelecanos, J., Slomka, S., Sridharan, S.: Enhancing automatic speaker identification using phoneme clustering and frame based parameter and frame size selection. *Fifth International Symposium on Signal Processing and Its Applications*, Vol. 2 (1999) pp. 633-636
66. Reynolds, D.A.: "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing* **2** (1994) 639{643,
67. Lareau, J.: Application of Shifted Delta Cepstral Features for GMM Language Identification. RIT, CS. Dept (2006)
68. Milner, B.: "A comparison of front-end configurations for robust speech recognition,". International Conference on ICASSP. Acoustics, Speech, and Signal Processing, IEEE (2002)
69. Openshaw, J.P., Sun, Z.P., Mason, J.S.: "A comparison of composite features under degraded speech in speaker recognition,". International Conference on ICASSP. Acoustics, Speech, and Signal Processing, IEEE (1993)
70. Lawrence, R., Hwang, J.B.-. "*Fundamentals of Speech Recognition*". Prentice Hall, New jersey (1993)
71. Bielefeld, B.: "*Language identification using shifted delta cepstrum*" Fourteenth Annual Speech Research Symposium (1994)
72. Allen, F.: *Automatic Language Identification*. University of New South Wales, Sydney (2005)
73. Lin, C.-Y., Wang, H.-C.: Language Identification Using Pitch Contour Information,. *Acoustics, Speech, and Signal Processing, ICASSP Vol. 1. IEEE International Conference* (2005) pp 601 - 604
74. Obuchi, Y., Sato, N.: Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization. *Acoustics, Speech, and Signal Processing, ICASSP Vol. 1. IEEE International Conference* (2005) pp 569 - 572
75. Cutler, A., Dahan, D., Donselaar-Van, W.: Prosody in the comprehension of spoken language: A literature review. *Language and Speech* **40** (1997) pp 141-202
76. Cummins, F., Gers, F., Schmidhuber, J.: Language Identification from Prosody without explicit Features. *EUROSPEECH'99*, Vol. 1, Budapest, Hungary, (1999) pp. 371-374
77. Samouelian, A.: "Automatic Language Identification using Inductive Inference". 4th International Conference on Spoken Language Processing ICSLP Philadelphia, USA. (1996)
78. **Huang, X., Acero, A., Hon, H.-W.**: *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Pearson Education (2001)
79. Modic, R., Lindberg, B., Petek, B.: Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English speechDat2". *NOLISP, Le Croisic, France* (2003)
80. Gupta, M., Gilbert, A.: "Robust speech recognition using wavelet coefficient features". *IEEE Automatic Speech Recognition and Understanding Workshop, USA* (2001) pp. 445-448
81. Murthy, H.A., Yegnanarayana, B.: Formant extraction from group delay function. *Speech Communication*, Vol. 10. Elsevier Science Publishers B. V. , Amsterdam, The Netherlands (1991) pp 209 - 221
82. Murthy, H.A., Gadde, V.: The modified group delay function and its application to phoneme recognition. *Acoustics, Speech, and Signal Processing ICASSP Vol. 1. IEEE International Conference* (2003) pp 68-71
83. Hegde, R.M., A.Murthy, H., Gadde, V.: Application of the Modified Group Delay Function to Speaker Identification and Discrimination. *ICASSP, Vol. 1* (2004) pp. 517-520

84. Hegde, R.M., A.Murthy, H., Gadde, V.: Continuous Speech Recognition using Joint Features derived from The Modified Group Delay Function and MFCC. INTERSPEECH-ICSLP (2004)
85. Database for Indian languages. Speech and Vision Lab, IIT Madras, Chennai, India (2001)
86. Hedge, R.M., Murthy, H.A.: Automatic Language Identification and discrimination using the modified group delay feature. *Conf. on Intelligent Sensing and Information Processing*, Chennai (2005) pp. 395-399
87. Greenberg, S., Arai, T.: What are the Essential Cues for Understanding Spoken Languages? *IEICE Transaction on Information & System* **E87-D** (2004) pp. 1059
88. Dimitriadis, V., Maragos, P., Potamianos, A.: Robust AM-FM Features for Speech Recognition. *IEEE Signal Processing Letters* **12** (2005) pp. 621-624
89. Dimitriadis, D., Maragos, P.: Robust Energy Demodulation Based on Continuous Models with Application to Speech Recognition.: *Eurospeech-03*, Geneva (2003)
90. Maragos, P., Kaiser, J.F., Quatieri, T.F.: Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing* **41** (1993) pp. 3024-3051
91. Potamianos, A., Maragos, P.: A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation. *Signal Processing Magazine, IEEE* **37** (1994) pp. 95-120
92. Thiruvaran, T., Ambikairajah, E., Epps, J.: Extraction of FM components from speech signals using all-pole model. *Electronics Letters* **44** pp. 449-450
93. Yin, T., Thiruvaran, T., Ambikairajah, E., Chen, F.: Introducing a FM based Feature to Hierarchical Language Identification. *InterSpeech 08* (2008) pp 731-734
94. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. *A Speaker Odyssey, The Speaker Recognition Workshop* (2001) pp 243-248
95. Torre, J., Segura, C., Benitez, A., Peinado, M., Rubio, A.J.: Non-Linear Transformations of the Feature Space for Robust Speech Recognition. *IEEE ICASSP*, Vol. 1 (2002) pp. 401-404
96. Allen, F., Ambikairajah, E., Epps, J.: Language Identification Using Warping and the Shifted Delta Cepstrum. *IEEE Int. Workshop on Multimedia Signal Processing*, Shanghai, China (2005)
97. Dempster, Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc* **39** (1977)
98. Reynolds, D.A., Quatieri, T.F., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Process.* **10** (2000) pp 19-41
99. Mat'ejka, P., Burget, L., Schwarz, P., Černocký, J.: Brno University of Technology System for NIST 2005 Language Recognition Evaluation. *IEEE 2006 Odyssey: The Speaker and Language Recognition Workshop* (2006)
100. Vapnik, V.N.: *Statistical Learning Theory*. Wiley, New York (1998)
101. Collobert, R., Bengio, S.: SVM-Torch: Support Vector Machines for Large-Scale Regression Problems. *J. Mach. Learn. Res.* **1** (2001) pp 143-160
102. Campbell, W.M.: Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. *The International Conference on Acoustics Speech and Signal Processing* (2002) pp. 161-164
103. Platt, J.C.: Probabilities for SV Machines. In: Smola, A.J., Bartlett, P.L., Schölkopf, B., Schuurmans, D. (eds.): *Advances in Large Margin Classifiers*, MIT, Cambridge (2000) pp. 61-74
104. Allen, F., Ambikairajah, E., Epps, J.: *Warped magnitude and phase-based features for language identification*. *ICASSP IEEE* (2006) pp. 201-204
105. Yan, Y., Barnard, E.: Experiments for an Approach to Language Identification with Conversational Telephone Speech. *International Conference on Acoustics Speech and Signal Processing* (1996) pp. 789-792
106. Zissman, M.A.: Predicting, Diagnosing and Improving Automatic Language Identification Performance. *Eurospeech'97* (1997) pp. 51- 54

107. Gauvain, J.L., Messaoudi, A., Schwenk, H.: Language Recognition Using Phoneme Lattices. International Conference on Spoken Language Processing (2004) pp. 2833-2836
108. Yin, B., Ambikairajah, E., Chen, F.: Hierarchical Language Identification based on Automatic Language Clustering. InterSpeech - EuroSpeech, Antwerp, Belgium (2007)
109. Yin, B., Ambikairajah, E., Chen, F.: Improvements on Hierarchical Language Identification based on Automatic Language Clustering. the IEEE International Conference on Acoustic, Audio and Signal Processing, Las Vegas, USA (2008)
110. Chu-Carroll, J., Carpenter, B.: Vector-based natural language call routing. *Computat. Linguist* **23** (1999) pp 361-388
111. Bellegarda, J.R.: Exploiting latent semantic information in statistical language modeling. *IEEE* **88** (2000) pp 1279-1296
112. Salton, G.: The SMART Retrieval System-Experiments in Automatic Document Processing Source Prentice-Hall, Inc. Upper Saddle River, NJ, USA (1971)
113. Kuo, H.K.J., Lee, C.-H.: Discriminative training of natural language call routers. *IEEE Trans. Speech Audio Process.* **11** (2003) pp 24-35
114. Gao, S., Wu, W., Lee, C.-H., Chua, T.-S.: A MFoM learning approach to robust multiclass multi-label text categorization. *ICML* (2004) pp. 42-49
115. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman Hall, New York (1984)
116. Haykin, S.: Neural Networks: A Comprehensive Foundation, McMillan, Englewood (1994)
117. Ma, B., Li, H., Lee, C.-H.: An Acoustic Segment Modeling Approach to Automatic Language Identification. INTERSPEECH-05 (2005)
118. Li, H., Ma, B., Lee, C.-H.: A Vector Space Modeling Approach to Spoken Language Identification. *IEEE Transactions on audio, speech, and language processing* **15** (2007)
119. Ma, B., Tong, R., Li, H.: Discriminative vector for spoken language recognition.: *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Vol. 4, Honolulu, Hi. (2007) pp 1001-1004

RT_015, noviembre 2009

Aprobado por el Consejo Científico CENATAV

Derechos Reservados © CENATAV 2009

Editor: Lic. Lucía González Bayona

Diseño de Portada: DCG Matilde Galindo Sánchez

RNPS No. 2142

ISSN 2072-6287

Indicaciones para los Autores:

Seguir la plantilla que aparece en www.cenatav.co.cu

C E N A T A V

7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;

Ciudad de La Habana. Cuba. C.P. 12200

Impreso en Cuba

