RNPS No. 2142
ISSN 2072-6287
Versión Digital

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

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Métodos de extracción de rasgos para la identificación del idioma: estado del arte

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

RT_036 **octubre 2010**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

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

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de extracción de rasgos para la identificación del idioma: estado del arte

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

RT_036 **octubre 2010**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Métodos de extracción de rasgos para la identificación del idioma: estado del arte

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_036 CENATAV
Fecha del camera ready: 30 de junio de 2010

Resumen. La identificación del idioma tiene como objetivo fundamental determinar el lenguaje en el que una persona está hablando, basándose en las características de su voz y el conocimiento previo del idioma, sin considerar al hablante y lo que está diciendo. Una de las fases más importantes en el flujo de actividades de un sistema de identificación automática de idioma es la concerniente a la extracción de rasgos. La definición de cuál es la fuente de información que se va a utilizar para la identificación del idioma y cuáles son las características que mejor lo representan son temas de especial interés en esta rama de investigación. Este trabajo ofrece un estado del arte de los métodos de extracción de rasgos que más se han utilizado en el área del reconocimiento del idioma.

Palabras clave: identificación de idioma, extracción de rasgos, parámetros articulatorios

Abstract. Language identification is essential to determine the target language in which a person is speaking, based on the characteristics of his voice and prior knowledge of the language, regardless of the speaker and what he is saying. One of the most important phases in the flow of activities of a system for automatic language identification is the extraction of speech features. The definition of what is the source of information to be used to identify the language and what are the characteristics that best represent it, are topics of particular interest in this field of research. This paper presents a state of the art of feature extraction methods used in the area of language recognition.

Keywords: Language Identification, Extraction of Features, Articulatory Parameters

1 Introducción

La identificación automática del lenguaje hablado (LID) es el proceso por el cual el idioma de una muestra de señal de voz digitalizada es reconocido por una computadora. En la actualidad existen un conjunto grande de sistemas cuyo resultado final depende en gran medida de la tecnología LID, constituyendo un módulo importante dentro de varias aplicaciones como sistemas de conversación multilingüe y traducción de lenguaje hablado, entre otros.

Las investigaciones para la identificación automática del lenguaje hablado se han desarrollado siguiendo diferentes enfoques. Sin embargo, todos ellos están estructurados de forma similar:

Fig. 1. Diagrama de bloques de un (*Sistema de identificación de idiomas)*

Tal como muestra la figura 1 el primer paso en la caracterización del lenguaje hablado es definir las fuentes de conocimiento o características que se van utilizar para distinguir un idioma de otro. Por lo tanto, un reto importante para estos sistemas es la incorporación efectiva a sus modelos de fuentes de conocimiento que sean discriminativas y robustas.

Algunos sistemas usan sólo las expresiones digitalizadas del discurso del habla 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 el proceso 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.

A continuación se aborda de forma general las diferentes formas de caracterización de los idiomas y los métodos de extracción de rasgos que más se han usado en los sistemas LID.

2 Identificación de los idiomas en los humanos

El proceso de identificación del lenguaje hablado tiene como objetivo fundamental determinar el idioma en el que una persona está hablando, basándose en las características de su voz y el conocimiento previo del mismo, sin considerar al hablante o lo que se está diciendo.

En estudios realizados se ha comprobado la capacidad que tienen los seres humanos para la identificación de los idiomas. Con escuchar la voz unos segundos, las personas son capaces de determinar de qué idioma se trata, siempre y cuando conozcan el idioma en particular; y en el caso de que sea un idioma con el que ellos no están familiarizados, pueden realizar un juicio subjetivo de acuerdo a los lenguajes similares que ellos conocen, por ejemplo, suelen decir: “suena parecido al alemán”. De acuerdo a esto, Muthusamy en 1994 [1] realizó un estudio para obtener los mejores indicadores que tienen los humanos en la identificación del lenguaje hablado. Sus pruebas consistieron en dos casos: la identificación del lenguaje hablado por personas monolingües, es decir, que sólo conocen su lengua materna; y el segundo caso por personas que conocen varios idiomas. Para el experimento se analizaron 28 personas (14 mujeres y 14 hombres); de los cuales fueron 10 hablantes nativos del lenguaje inglés y 2 personas para cada uno de los 9 lenguajes restantes. Todas las personas conocían el lenguaje inglés. Las personas podían escuchar 10 segundos de señal de voz espontánea, del corpus OGI_TS [2], tantas veces como ellos desearan.

Después de dos o tres días las personas tenían que identificar el lenguaje de una muestra específica. Los resultados se muestran en las figura 2.

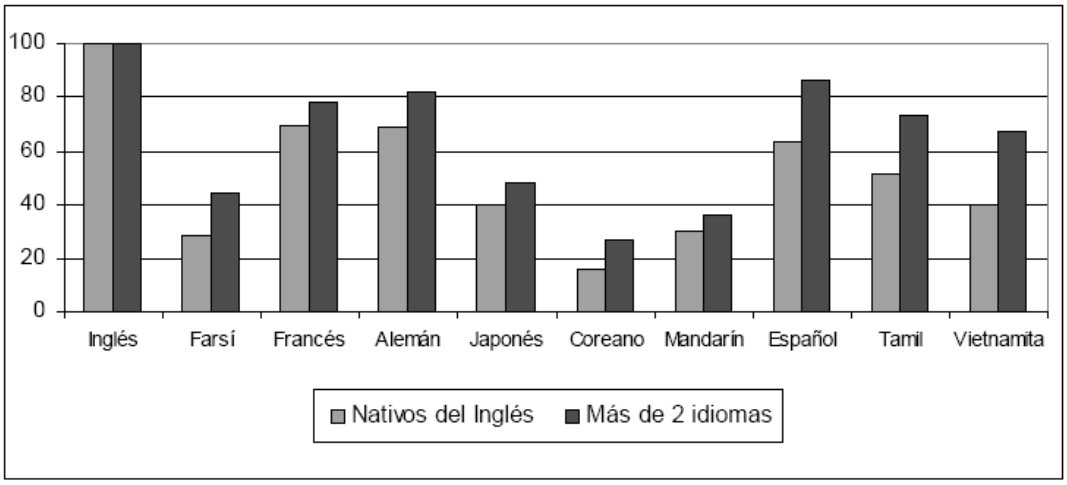

Fig. 2. Porcentaje de identificación del lenguaje hablado por hablantes (nativos del inglés) y los que hablan (más de dos idiomas) en 6 segundos de señal de voz (tomada de [1])

Como resultado de este experimento se obtuvo que el porcentaje de reconocimiento para personas que conocían 4 idiomas fue de 66.7%, para personas que conocían 3 idiomas 57.9%, para las personas que conocían 2 fue de 51.1% y las que sólo conocían uno fue de 44.1%. Con lo cual se concluyó que el porcentaje de discriminación de los idiomas aumenta cuando se tiene conocimiento de más lenguas. Lo importante de esto es notar que incluso la identificación del lenguaje hablado hecho por los humanos no tiene grandes porcentajes de discriminación, aún en el caso de personas que dominan cuatro idiomas. [3]

Ahora bien, de acuerdo a lo antes expuesto, la identificación automática del lenguaje hablado es el proceso por el cual el idioma de una muestra de señal de voz digitalizada es reconocido por una computadora. Las investigaciones realizadas han revelado la factibilidad de este proceso. Partiendo del análisis de una señal de voz se puede determinar el idioma en que se está hablando gracias a la existencia de determinados parámetros característicos de este tipo de señales cuya información resulta significativa en el proceso de identificación y reconocimiento de los idiomas.

3 Producción y percepción del habla

3.1 La producción del habla: el aparato fonador

El aparato fonador es el conjunto de órganos que se encargan de la emisión del sonido. Su funcionamiento está controlado por el sistema nervioso central. La figura 3 ilustra las partes en que se divide y los órganos que se agrupan en cada una.

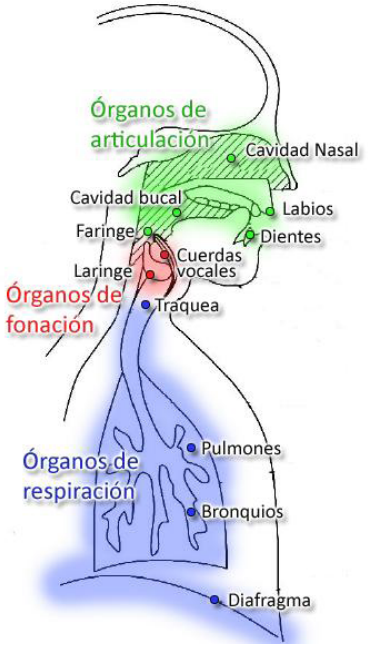

Fig. 3. Aparato fonador

El sonido que conforma al habla se obtiene como resultado de la circulación de aire por el aparato fonador. El proceso se inicia con la expiración de aire procedente de los pulmones que circula pasando por los bronquios y la tráquea hasta llegar a la laringe. Aquí se encuentra con las cuerdas vocales que son una serie de repliegues o labios que al vibrar producen sonidos.

Hay cuatro cuerdas vocales, dos superiores y dos inferiores. Estas últimas son las responsables de la producción de la voz. Se puede decir que son dos pequeños músculos elásticos que al dejar pasar el aire libremente, sin hacer presión, se produce el proceso de respiración. Ahora, si se juntan, el aire choca contra ellas, produciendo el sonido conocido como voz. La frecuencia a la que vibran las cuerdas vocales se le llama frecuencia fundamental. [4]

El aire al sobrepasar las cuerdas vocales atraviesa la cavidad bucal, que no es más que un resonador o filtro acústico y su función es responder selectivamente a las frecuencias de vibración que coincidan con la suya. La cavidad bucal da lugar a la producción de resonancias cuando la señal acústica que lo atraviesa tiene componentes de frecuencia coincidentes con las suyas. De estas resonancias solo se consideran las tres o cuatro primeras, que son los llamados 'formantes' y cubren un rango de frecuencias entre 100 y 3500 Hz., ellos concentran la mayor parte de la información existente en la señal. Este comportamiento se debe a que la característica transferencial del tracto vocal es similar a la de un filtro pasabajo con una caída de aproximadamente -12 dB por octava, y por lo tanto atenúa las resonancias de alta frecuencia. La cavidad bucal está controlada por los órganos articulatorios (lengua, dientes, labios) que al cambiar de posición conforman cavidades de volumen o resonadores de formas diferentes, lo que ocasiona formas de onda diferentes y por tanto espectros diferentes a la salida del resonador [5]. La cavidad nasal complementa al tracto vocal en el caso de las consonantes nasales.

3.2 La percepción del habla

El receptor del ser humano se conforma por el sistema oído-cerebro. Ellos se encargan de captar las ondas acústicas y procesarlas. El oído humano (fig. 4) está formado por tres secciones: el oído externo, el oído intermedio y el oído interno.

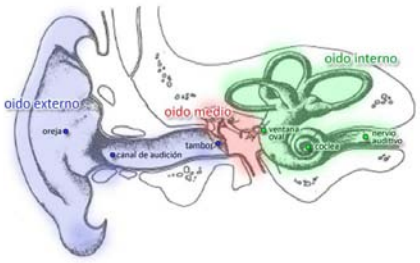

Fig. 4. Estructura del sistema periférico de audición

El oído externo está formado por el músculo externo que se conoce por oreja y el canal de audición externo. Las ondas sonoras entran por la oreja, se amplifican y se encaminan hacia el interior. El conducto auditivo es un tubo de 2,5 cm. de largo aproximadamente que finaliza en un recubrimiento que se llama tambor.

Las variaciones de presión provocadas por la señal sonora que alcanzan al tambor ocasionan que este entre en resonancia. Por lo que el tambor va a vibrar a la frecuencia de la onda sonora que lo excitó. Esta vibración se transmite hacia el oído interno pasando por el oído intermedio, el que además de ocuparse de la transmisión del sonido hacia el oído interno, se encarga de ajustar el valor de la intensidad del sonido para que sea suficientemente alta para el procesamiento y baja para que no afecte a los órganos que componen al oído.

En el oído interno se encuentra la cóclea que es un tubo espiral de 3,5 cm. de longitud. Ella se encarga de detectar las frecuencias de la onda sonora, convertirlas en impulsos nerviosos y mandarlas al cerebro. La cóclea hace función de banco de filtros cuyas salidas se utilizan para estimar la ubicación de la fuente que emite la señal. Los filtros cercanos a la base de la cóclea responden a las altas frecuencias y los cercanos a su parte superior responden a las bajas frecuencias [6].

El oído no percibe todas las frecuencias de las señales sonoras con la misma sensibilidad. Las frecuencias altas y bajas se escuchan con menor intensidad. La zona de mayor fuerza está alrededor de los 3 kHz. La impresión de variación de frecuencias sigue una escala logarítmica, lo que implica que para notar la misma diferencia entre varias frecuencias es necesario duplicar su valor [4].

4 Niveles de información 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 es la incorporación efectiva a sus modelos de fuentes de conocimiento que sean discriminativas. Algunos sistemas usan sólo las expresiones digitalizadas del discurso del habla 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 en ocasiones muy caro computacionalmente.

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 [7].

4.1 Información acústica

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 segmentos de tiempo 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 representar la señal digitalizada de lenguaje como un vector de características.

Teniendo los mismos 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 [8] y enfoques que también utilizan los patrones de cambio de estos vectores de características con el paso del tiempo [9].

Siguiendo esta línea se ha investigado una variada colección de modelos computacionales que incluyen los modelos de mezclas gaussianas GMM [10], los modelos ocultos de Markov [10, 11], las redes neurales artificiales ANN [12] , y las máquinas de vectores de soporte SVM [13]. 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 son tan exactos como los sistemas que usan conocimiento lingüístico; sin embargo, pueden ser logradas mejoras en la exactitud integrando fuentes de conocimiento acústico y de bases lingüísticas [14, 15].

4.2 Información fonológica

La fonología es una rama de la lingüística que estudia los sistemas sonoros de los lenguajes humanos [16]. 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, tal 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, se ha comprobado que son una abstracción útil para el procesamiento de voz. Ladefoged [17] supone que hay probablemente acerca de 600 consonantes diferentes provenientes de los lenguajes del 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 y el más comúnmente usado por los lingüistas es el Alfabeto Fonético Internacional (IPA)[18]. Este alfabeto 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 todos los fonemas expresados en todos los lenguajes humanos, cada símbolo sirve para distinguir entre 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 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 once (para Rotokas, un lenguaje papú)[19] hasta por encima de cien (en ciertos lenguajes Khoisan del sur de África)[20]. Es probable que ninguno de los lenguajes comparta un inventario de fonemas exactamente igual. Además, aun si los dos lenguajes compartieran un conjunto común de fonemas, los fonemas diferirán en sus patrones de frecuencia relativa. [12]

Inventario de categorías fonéticas: Los patrones de categorías fonéticas amplias (ejemplo: vocales, consonantes fricativas, explosivas, nasales, y líquidas) también han sido utilizados para distinguir entre lenguajes en un intento para evitar la necesidad del reconocimiento fonético fino. Muthusamy [12] evaluó el uso de siete categorías fonéticas amplias - vocal, consonante fricativa, parada, cierre o silencio, sonido sonoro prevocálico, sonido sonoro intervocálico, y sonido sonoro postvocálico - Sin embargo, Hazen [14] encontró que, aun cuando los reconocedores de categorías fonéticas tienden a ser más precisos que los reconocedores fonéticos tradicionales que son más minuciosos y finos, si se usan pocos inventarios de categorías muy amplios, los sistemas de identificación de lenguaje tienen menor exactitud.

Rasgos Fono-tácticos: Los rasgos fono-tácticos 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 fono-tácticos. Los primeros en proponer el uso de estos rasgos fueron House y Neuburg [21], 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 secuencias 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[14]. 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 rasgos fono-tácticos. Los LID basados en fonemas usan un solo reconocedor fonético entrenado con algún idioma arbitrario con suficientes recursos (no necesariamente uno de los lenguajes a identificar) para tokenizar la entrada de datos hablados en los sonidos fonéticos para ese lenguaje, y después usan los modelos probabilísticos de 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 [22].

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[22].

Una tercera variante sería entrenar un reconocedor fonético con una base de datos fonética de amplia cobertura[17]. Para reducir el tiempo y el costo para desarrollar sistemas hablados en un nuevo idioma, los investigadores han estado trabajando en el desarrollo y el uso de un conjunto de sonidos fonéticos multilingües que representen los sonidos de los idiomas que van a ser modelados. Schultz y Waibel [23] en su investigación sobre el reconocimiento del lenguaje hablado multilingüe definieron un conjunto de sonidos fonéticos que cubren 12 lenguajes. Ellos suponen que las representaciones articulatorias de los fonemas son tan similares a través de los diferentes lenguajes, que los fonemas pueden ser considerados como unidades que son independientes del idioma subyacente[24].

Hazen y Zue [14, 15] desarrollaron un sistema LID que se basó en 87 unidades fonéticas independientes del idioma que fueron obtenidas agrupando a mano aproximadamente 900 etiquetas de sonidos fonéticos encontradas en las transcripciones. Los investigadores del Laboratorio LIMSI también han estado investigando el uso de un conjunto de sonidos fonéticos universales [25, 26] y han tratado de identificar criterios acústicos objetivos para agrupar los sonidos fonéticos dependientes en lenguaje. Corredor- Ardoy [27], encontraron que su sistema de LID usando un conjunto de sonidos fonéticos independiente del lenguaje se comportó tan bien como los mejores métodos empleando sonidos fonéticos dependientes del lenguaje. Ma y Li [28] 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 sonidos 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 usadas en 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 [14]. Kirchhoff y Parandekar [29] utilizaron un conjunto de categorías seudo-articulatorias que fueron diseñadas para captar las características del proceso de producción del habla, que incluyen: la manera de articular, el lugar consonántico de la articulación, el lugar vocálico de la articulación, el redondeando de los labios, la posición de la lengua (trasera o delantera), la sonorización, y la nasalidad. Ellos desarrollaron un método alternativo para la identificación de lenguaje 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.

4.3 Información silábica

Una sílaba es una unidad de pronunciación que generalmente es mayor que un fonema, 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) [16]. Los lenguajes pueden ser caracterizados por tipos comunes de sílabas, definidos en términos de secuencias de consonantes (C) y vocales (V) [30, 31] y generalmente permiten o prohíben ciertos tipos de estructuras de sílabas; esta representación puede ayudar a

discriminar entre los lenguajes. Por ejemplo CCCCVC es un tipo válido de sílaba en ruso, pero no en la mayoría de otros lenguajes.

Zhu et al. [32] desarrollaron sistema LID cuyo decodificador acústico produce un flujo de sílabas que fue usado en modelos de lenguaje de n-gramas basados en sílabas (en vez de estar basados en sonidos fonéticos).

El acento de los interlocutores extranjeros del idioma inglés se manifiesta diferente dada su localización dentro de la sílaba, hecho que se ha usado para mejorar la identificación del acento en este lenguaje [33].

4.4 Información prosódica

La duración, el tono, y los patrones de énfasis en un lenguaje a menudo difieren de otro lenguaje, cada lenguaje tiene patrones bien definidos de entonación. En los lenguajes de énfasis, el tono es a menudo usado para señalar la prominencia de la sílaba en la palabra mientras que en los lenguajes de tono, un cambio en el significado de una palabra es señalado por el tono en las sílabas (el chino-mandarín o el tailandés). Para los lenguajes de énfasis, los patrones de énfasis pueden proveer una indicación 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 indicaciones prosódicas de duración también pueden ser potencialmente útiles, por ejemplo, algunos lenguajes como el finés, hacen distinción de vocales y/o consonantes breves y largas.

Una información prosódica notablemente útil, investigada por lingüistas y psicólogos es el ritmo de los lenguajes, existiendo diferencias entre lenguajes en los que se midió el tiempo de énfasis (en el que las sílabas acentuadas resultaron ser más largas que las sílabas sin acentuar, como por ejemplo el inglés), el tiempo de cada en sílaba (en el que cada sílaba tiene duración comparable, como por ejemplo el francés), y el tiempo vocálico de las silabas (en el que cada tiempo vocálico tiene duración esencialmente constante, como por ejemplo el japonés)[34]. Aunque esta clasificación por el ritmo permanece en constante discusión [35], el modelado rítmico fue investigado por Rouas et al. [36, 37] 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 a partir de un discurso espontáneo [36]. Ciertamente, extraer información prosódica fidedigna de un lenguaje, 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 generalmente son combinadas con otras fuentes para lograr una exactitud razonable. Muthusamy [12] fue capaz de incorporar la variación del tono, la duración, y las características de proporciones entre sílabas en su modelo para LID. Hazen y Zue [14, 15] 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. [38] integró exitosamente características prosódicas (duración y tono), del espectro, fonotácticas, y la bolsa de sonidos, donde la variación del tono y la duración de los fonemas fueron especialmente útiles para identificar el lenguaje con pequeños segmentos de discurso hablado.

4.5 Información léxica

Cada lenguaje tiene su propio vocabulario lexicológico, lo cual debe ayudar a identificar un lenguaje más confiadamente, pero esto requiere de un sistema de reconocimiento del lenguaje

hablado para cada uno de los lenguajes candidatos, con el entrenamiento requerido, para asegurar que el modelo lexicológico del lenguaje es el adecuado. Schultz et al. [39, 40] desarrollaron un sistema LID para cuatro lenguajes basándose en el amplio léxico del reconocimiento del habla continua (LVCSR). Ellos encontraron que los sistemas basados en palabras con modelos tri-gramas de palabras funcionan significativamente mejor que los sistemas basados en fonemas con modelos tri-gramas de fonemas, para la identificación de los cuatro lenguajes, sugiriendo que el nivel léxico provee mayor habilidad para discriminar entre los lenguajes, aún cuando las tasas de errores de palabra sean medianamente altas. Matrouf et al. [41] encontraron que incorporando además información léxica a un enfoque basado en fonemas se produjeron disminuciones relativas del error del 15–30%, y este alcance incremental del léxico para cada lenguaje tuvo un efecto positivo sobre el funcionamiento del sistema. Hieronymus and Kadambe [42] 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 de prueba respectivamente, sin usar medidas de confianza, y 93% y 98% de identificaciones correctas con medidas de confianza. Aunque cada lenguaje obviamente tiene su vocabulario lexicológico 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.

4.6 Información morfológica

La morfología es una rama de la gramática que investiga la estructura de las palabras [16]. El campo de la morfología está dividido en dos sub-campos: La morfología aglutinativa, que investiga los afijos (prefijos y sufijos) 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 usada para crear una nueva palabra con una categoría gramatical diferente, como el sustantivo modificación deducido del verbo modificar. Como cada lenguaje forma palabras siguiendo diferentes formas, la morfología podría proveer una indicación excelente para identificación automática de los mismos. Por ejemplo, pueden emplearse sufijos comunes para discriminar entre algunas de las lenguas romance ("ment" en francés, "miento" en español, "mento" en portugués y "mente" en italiano). Aunque en el lenguaje la morfología de una palabra es poco notoria, debido en parte a la predominancia de la información fono-táctica, considerando el conocimiento morfológico de los lenguajes a identificar, un sistema LID podría enfocarse sobre porciones específicas de palabras para discriminar entre dos lenguajes.

4.7 Información sintáctica

Los lenguajes y los dialectos también difieren en las formas en que las palabras son organizadas 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 como los acentos, los errores en el uso gramatical podrían proveer una indicación 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 muy probable que las palabras aparezcan en contextos diferentes a través de las frases, para diferentes lenguajes. Aunque las palabras pueden ser suficientes como para discriminar entre lenguajes, la sintaxis 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 sintaxis más directamente usando modelos estructurados de lenguaje [43, 44].

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, existe una correspondencia entre la exactitud (eficacia) y la eficiencia, y además, está la necesidad de poseer los recursos suficientes para sustentar el conocimiento introducido en el sistema automático. En este sentido se ha llegado a un balance de ingeniería; usar aquellos recursos más 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 de habla disponible) afectará las fuentes de conocimiento que pueden utilizarse confiadamente para determinar su lenguaje. Mientras más pequeñas sean las muestras usadas, menos probable es que un volumen significativo de conocimiento de alto nivel esté disponible para discriminar un lenguaje particular de los demás.

5 Parámetros que caracterizan a la señal de voz

5.1 Parámetros acústicos y perceptuales

Las señales sonoras y en especial las de voz, tienen propiedades físicas que las caracterizan. Estas propiedades son cuantificables por lo que se han definido una serie de magnitudes para medir su valor. Por otro lado el oído aprecia estas características, por lo que se han definido una serie de atributos perceptuales. Existe una relación de correspondencia predominante entre un atributo perceptual y una propiedad física específica. Pero no se puede afirmar que sea exactamente unívoca. A continuación se muestra la relación entre las magnitudes físicas que caracterizan al sonido y su correspondiente atributo perceptual [5, 6]:

- Intensidad ► Volumen
  La intensidad es la potencia acústica con que se expulsa el aire de los pulmones. Se mide en dB y depende de la amplitud de la señal de voz. El volumen se sonido es la impresión de alto o bajo correspondientes a los dB de potencia.
- Forma espectral ► Timbre o sonoridad

La forma espectral es la forma que toma la señal de voz en el dominio de la frecuencia. Es la distribución de los armónicos con sus respectivas amplitudes en un eje de amplitud vs. frecuencia. El timbre o sonoridad es el resultado del número y organización de los armónicos. Es grave cuando los armónicos de mayor amplitud están en las bajas frecuencias, y agudo cuando hay concentración de armónicos de gran amplitud en las frecuencias altas.

- Frecuencia de resonancia ► Tono
  La frecuencia de resonancia es la frecuencia natural de oscilaciones de las cuerdas vocales. También se denomina frecuencia o armónico fundamental. El tono es la impresión auditiva que se percibe de la frecuencia fundamental, se sitúa en una escala de altos y bajos. Su función en la oración es la entonación.

A continuación se abunda en otras características de la señal de voz.

- Formantes
  Son los máximos de intensidad en el espectro de un sonido correspondientes a una frecuencia determinada. Sirven para distinguir componentes del habla humana, sobre todo las vocales y sonidos sonoros. Se encuentran en la zona del espectro de máxima intensidad, por lo que alrededor de las frecuencias de los formantes aparecen grandes concentraciones de energía. De la relación entre los formantes depende el timbre del sonido.
  En el caso de la señal de voz coinciden con las resonancias del conducto vocal. En algunos casos pueden coincidir con la frecuencia de otros armónicos de la señal de voz. Los formantes se nombran F1, F2, F3, F4, etc. en orden creciente de frecuencia. Es decir la f(F1)< f(F2)< f(F3)< f(F4). A medida que transcurre el tiempo el valor de intensidad de cada formante varía. Generalmente los dos primeros formantes se utilizan para caracterizar una vocal.
  Investigaciones previas (8) han demostrado que la intensidad de los formantes y de la frecuencia fundamental varía con la edad y el sexo. Los valores de la frecuencia fundamental varían de forma inversamente proporcional al incremento de edad, sin embargo, no se aprecian diferencias significativas de este parámetro atendiendo al sexo. En el caso de la intensidad de F1, F2 y F3, los valores se hacen menores a medida que aumenta la edad, y las componentes de frecuencias de las niñas son superiores a las de los niños.
- Duración-Velocidad
  La duración es el tiempo de extensión de un fonema. La velocidad es el resultado de las extensiones de los fonemas alineados en segmentos (palabras y frases, incluidas las pausas). La velocidad también está relacionada con la fluidez con que se expresan las ideas.
- Monotonía
  Aunque su nombre alude sólo al tono, se produce al hablar manteniendo fijos el tono, el volumen y la velocidad. La monotonía dificulta al oyente el entendimiento, por la confusión de unas palabras con otras (suenan todas igual). Algunos estudios han demostrado que la monotonía disminuye la comprensión en más del diez por ciento.
- Ritmo
  Es una cadencia particular de la locución que la hace armónica. Se habla de ritmo cuando es posible prever lo que va a seguir en función de lo percibido. Esa previsibilidad da la oportunidad de entrar en sintonía con la fuente emisora y seguir la cadencia con el cuerpo (bailando) o la mente (seguir la letra de una canción coincidiendo en tiempo con el cantante). Puede ser sostenido o irregular. El ritmo sostenido es más agradable, en

función de su musicalidad. Pero conlleva el riesgo de caer en el canto. El ritmo irregular se asocia con ciertos estados de ánimos, lo que lo hace un parámetro importante de la retórica.

En general, el ritmo y la tonalidad, expresan las tensiones afectivas y emocionales y son percibidos mucho antes que su contenido propiamente semántico. La musicalidad de la locución es el producto de la distribución armoniosa de los sonidos en el tiempo[45].

5.2 Los parámetros lingüísticos en la identificación del idioma

Desde el punto de vista lingüístico, los antecedentes en la identificación del idioma muestran que los especialistas en esta área han intentado realizar la clasificación de los lenguajes humanos basados en características suprasegmentales, que actúan simultáneamente sobre más de un segmento (al menos sobre la sílaba): el acento, el tono (o la sucesión de ellos, es decir, la entonación) y la duración. El conjunto de estos elementos suprasegmentales se denomina prosodia, que incluye la entonación, los patrones de acentuación, el ritmo, la melodía, etc. y en la realización de los mismos intervienen los siguientes índices acústicos y articulatorios:

1. La vibración de las cuerdas vocales que es la fuente de sonoridad de los segmentos sonoros, y también del movimiento del tono fundamental, que puede utilizarse en la distinción de las palabras (tono) o de oraciones (entonación).
2. Todo segmento tiene una dimensión temporal, es decir, una duración. Ésta, además, puede desempeñar, en determinadas lenguas, una función distintiva.
3. Todo segmento, al realizarse, ha de tener alguna intensidad. Esta, además, puede desempeñar en algunas lenguas una función distintiva (acento) [3].

A continuación se describen los rasgos suprasegmentales que constituyen la prosodia.

- El acento

El acento recae sobre una sílaba de la cadena hablada y la destaca o realza frente a otras no acentuadas [5].

Esta prominencia silábica suele interpretarse tradicionalmente como reflejo de intensidad; por eso, se le ha llamado "acento de intensidad". La realidad, sin embargo, es más compleja: la prominencia resulta de la conjunción de varios factores articulatorios:

1. Una mayor fuerza respiratoria, que genera una mayor intensidad.
2. Una mayor tensión de las cuerdas vocales, que genera una elevación del tono fundamental.
3. Una mayor prolongación en la articulación de los sonidos, que supone un aumento de la duración silábica.

Así pues, la sílaba acentuada, habitualmente, es más intensa, más alta y más larga que las sílabas adyacentes.

En cuanto a la posición que la sílaba acentuada ocupa dentro de la frase, algunas lenguas son de acento libre, es decir, no hay manera de prever en qué sílaba recae el acento (inglés); otras, por el contrario, son de acento fijo, es decir, la posición del acento es siempre previsible (francés).

En las distintas lenguas del mundo, el acento puede tener las siguientes funciones lingüísticas:

1. Contrastiva: distingue la sílaba acentuada: "El libro es de él".
2. Distintiva: distingue unidades en lenguas con acento libre: "amo"/"amó".
3. Demarcativo: en lenguas de acento fijo, señala los límites de las unidades en la secuencia. (el final de una palabra en turco).

4. Culminativa: en las lenguas de acento libre, señala la presencia de una unidad acentual, sin indicar sus límites.

- La entonación

La entonación es uno de los componentes más complejos de una lengua. Se ha definido de muchas maneras: por el tono fundamental, por una conjunción de parámetros acústicos (tono, acento y duración, primordialmente), por su función lingüística, etc.

Quilis [5] define la entonación como "la función lingüísticamente significativa, socialmente representativa e individualmente expresiva de la frecuencia fundamental en el nivel de la oración".

Desde el punto de vista articulatorio, el tono depende básicamente de las cuerdas vocales: de su longitud, su grosor y su tensión pero, además, hay una serie de factores fonéticos que la condicionan:

1. Existe una relación entre la cualidad o el timbre de la vocal y la altura relativa de su frecuencia fundamental, de modo que las vocales más altas tienen un tono fundamental más elevado.
2. Las frecuencias fundamentales más altas aparecen después de las consonantes sordas, y las más bajas, tras las consonantes sonoras.
3. Además del tono fundamental, la duración y la intensidad también intervienen en la producción y la percepción de la entonación.

- La duración

La duración es también un fenómeno suprasegmental, puesto que cada sonido posee una duración propia. Así por ejemplo, es sabido que las consonantes fricativas son más largas que las oclusivas, que las sordas son las más largas que las sonoras, etc.

Algunas lenguas poseen pares de fonemas en función de la duración. Por ejemplo, el italiano distingue entre ciertas consonantes breves y largas o "dobles". El latín clásico distinguía entre vocales breves y largas.

De acuerdo a la articulación, la duración se basa en el mantenimiento por más o menos tiempo de una determinada configuración articulatoria. Basándose el concepto de coarticulación en que los sonidos no se pronuncian aislados, la proximidad articulatoria de unos con otros hace que se influyan mutuamente.

Ahora bien, dentro de la cadena hablada, los segmentos se agrupan en unidades cada vez mayores: sílabas, palabras y enunciados.

El núcleo silábico es la parte central de la sílaba, la que tiene mayor intensidad sonora y que se manifiesta en un espectrograma con una mayor amplitud. Desde el punto de vista acústico, los fonemas situados antes del núcleo silábico presentan un aumento de su intensidad, sonoridad y perceptibilidad, hasta llegar al máximo que constituye el núcleo. De igual manera, los fonemas que se encuentran detrás del núcleo presentan una disminución de dichas características, a partir del máximo constituido por el núcleo.

Desde el punto de vista articulatorio, los fonemas anteriores al núcleo silábico experimentan una abertura gradual de los órganos articulatorios, hasta llegar al máximo del núcleo, a partir del cual los fonemas experimentan un cierre. Lo mismo cabe señalar de la tensión articulatoria y de la presión del aire aspirado.

De acuerdo a lo anterior, la frontera o el límite silábico ha de estar situado donde se produce un mínimo entre dos máximos (es decir, los núcleos de las dos sílabas entre las que se establece el límite):

Los mínimos y máximos, como se ha dicho, corresponden a la intensidad, a la sonoridad, a la presión respiratoria, a la tensión muscular e, incluso, a la energía articulatoria general.

- El ritmo

El "ritmo" es probablemente el rasgo de la base articulatoria de un idioma cuya adquisición o dominio resulte más difícil. Según la lingüística, este término puede tener, al menos, dos acepciones:

1. En un sentido amplio se llama ritmo a las sensaciones auditivas que se perciben a los intervalos regulares de tiempo, producidas por repeticiones isofónicas de cualquier recurso prosódico del lenguaje, como puede ser la rima, la censura, etc.
2. En un sentido estricto, el ritmo es un rasgo básico de la cadena hablada, junto con la entonación y el acento. Aún siendo conscientes por una parte, de que lo que realmente se percibe auditivamente es una prominencia, conviene separar, en lo posible, los rasgos de tensión y los de melodía que se manifiestan en la cadena hablada; los rasgos de melodía corresponden a la entonación y los rasgos de tensión corresponden al ritmo (también llamado ritmo verbal para diferenciarlo del ritmo musical).

El ritmo de un grupo fónico es la pauta de tensión formada en el mismo por la combinación de sílabas acentuadas y no, largas y breves. El ritmo es uno de los rasgos más característicos de un idioma. Como no todos los idiomas hacen el mismo uso de las sílabas largas y breves, y de las acentuadas o no, habrá distintos tipos de ritmos; los más importantes son el acentual y el silábico.

Ritmo acentual (o stress-timed) quiere decir que las pautas que se forman en el grupo fónico tienen un tempo marcado por el acento, o sea, están acompasadas por el acento, mientras que en el ritmo silábico (o syllable timed) el ritmo está acompasado por la sílaba.

5.3 Los parámetros articulatorios en la identificación del idioma

En la mayoría de los sistemas de identificación de idiomas, la representación acústica de la señal de voz provee información referida a la distribución de la energía de la señal a través del tiempo y de la frecuencia, y esta representación es usada como fuente de información sobre la secuencia de las palabras. Recientemente se han ido incorporando otras fuentes de información alternativas tales como las características prosódicas que reflejan la información lingüística de la señal, como se explico en el epígrafe anterior.

Teniendo en cuenta esta tendencia se propone entonces hacer un análisis de la información que podrían aportar los parámetros articulatorios a esta tarea, ya que investigaciones realizadas plantean que los parámetros articulatorios describen las propiedades de la producción del habla mucho mejor que las propiedades acústicas.

Pero, ¿tienen alguna relación las características articulatorias en el proceso de percepción humana del habla?

Durante las décadas correspondientes a 1950 y hasta 1970, se desarrollaron un buen número de estudios de desorientación de percepción, lo cual a menudo ha sido citado como una fuente de evidencias que prueban la realidad perceptual de las características articulatorias/acústicas – fonéticas. Estos estudios [46-48] han demostrado que las confusiones de percepción de vocales y consonantes bajo condiciones de escucha limpia y ruidosa a menudo se modelan a lo largo de las dimensiones articulatorias: Los segmentos que son altamente confundibles en su mayor parte pueden ser descritos como que difieren en sólo una o dos características articulatorias.

Miller y Nicely [46], por ejemplo, estudiaron la confusión de consonantes iniciales delante de /a/ en sílabas sin sentido dentro de un discurso de habla limpio, el discurso filtrado, y el discurso con ruido blanco aditivo, con varias relaciones señal/ruidos. Fueron producidas 200 sílabas por cinco locutores diferentes y transcritas por cuatro oyentes diferentes. Para cada

condición acústica, las matrices de confusión de los que emitieron vs. los que percibieron las consonantes fueron calculadas y analizadas.

Mientras que el filtrado paso alto del habla distribuyó errores de percepción que no mostraron ningún patrón fonético, el filtrado paso bajo en su mayor parte produjo confusiones entre segmentos que fueron similares en términos de los parámetros articulatorios. Todas las consonantes fueron entonces agrupadas en superclases definidas por cinco dimensiones articulatorias: sonorización, nasalidad, fricación, duración y lugar. Los autores llegaron a la conclusión de que los cinco grupos de características fueron perceptivamente independientes y propusieron un modelo de percepción de discurso multi-canal, que asume la existencia de canales de percepción independientes asociados a la interpretación de los parámetros articulatorios, la salida del cual se combinan para identificar clases fonéticas.

Note que ese experimento no prueba la existencia de cinco canales de percepción que se corresponden precisamente a los anteriormente citados grupos de características. La interpretación de sus datos es hasta cierto punto predeterminada por una elección a priori de las cinco categorías descriptivas. No hay prueba que desmienta la existencia de menos o más de cinco canales de percepción. Aunque las características articulatorias descritas en este estudio fueron descubiertas de forma concisa y naturalmente, nada sugiere que los oyentes humanos las exploten activamente en el proceso de percepción de discurso.

Una fuente adicional de prueba a favor de la realidad de percepción de características articulatorias proviene de experimentos de similitud de juicio. En estos estudios fue encontrado que los sonidos fueron juzgados más similares por oyentes humanos cuando tuvieron más características articulatorias / fonéticas en común. [49] [50]

Por otra parte existen teorías fonéticas y físico-lingüísticas que toman un punto de vista extremo sobre la pregunta de la relevancia de percepción de las categorías articulatorias y reclaman que las categorías articulatorias forman exclusivamente la base para la percepción humana del habla, o sea que los oyentes perciben los sonidos reconstruyendo los gestos articulatorios pertinentes a partir de la señal de habla. La primera teoría de esta clase fue la Teoría Motora de Percepción de Discurso, de Liberman en 1967 [51] y revisada por Liberman y Mattingly en 1986 [52], son también teorías desarrolladas la “Modularity Theory” de Fowler y Browman [53] y la “Articulatory Phonology” de Goldstein [54].

Los proponentes de dichas teorías sostienen la opinión de que es más verosímil suponer que los humanos poseen una sola representación y / o módulo de procesamiento para la producción y la percepción de discurso, que postular la existencia de dos módulos separados, altamente especializados para estas tareas. Proponen que la producción y la percepción comparten una representación común y quizá una estrategia común de procesamiento, todo lo cual es probablemente innato.

La teoría motora cuenta con el respaldo de ciertas conclusiones experimentales y ciertas observaciones empíricas, como el hecho que los oyentes humanos, en situaciones donde encuentran problemas de percepción de discurso, involuntariamente pronuncian los movimientos articulatorios correspondientes, pero por otra parte, hay oyentes que son incapaces de articular sonidos fonéticos (mudos), pero pueden comprender discursos. Además, aunque los gestos articulatorios son contextualmente menos variables que los parámetros acústicos, necesitan ser rescatados de la señal acústica, lo cual requiere una transformación complicada semejante a la percepción de discurso puramente auditiva. En resumen, resulta altamente dudoso que los gestos articulatorios sean indispensables para la percepción humana de sonidos fonéticos. Sin embargo, parece que probablemente en ciertas situaciones las representaciones articulatorias pueden ayudar a clasificar sonidos fonéticos dentro de las categorías lingüísticas.

Las características articulatorias son la representación de algunas importantes propiedades fonológicas que aparecen durante la producción del habla y pueden considerarse clases abstractas que describen los movimientos o posiciones de los diferentes articuladores durante la producción del habla. Generalmente estas clases caracterizan los aspectos más esenciales de la articulación en una forma altamente discretizada y canónica, conduciendo a un nivel figurativo intermedio entre la señal acústica y las unidades léxicas. Las características articulatorias han sido usadas como rasgos en las reconocimiento del habla [55, 56], reconocimiento del locutor [57, 58] e identificación de idiomas [59, 60]. En la tabla 1 se muestran la características articulatorias más utilizadas y los posibles valores que pueden tomar.

Tabla 1. Características articulatorias.

Características articulatorias	**Valores que toman**
Voicing (sonoridad)	voiced , voiceless (sonoro, no sonoro)
Manner (manera)	vowel, nasal, lateral, fricative, stop (Vocal, nasal, lateral, consonante fricativa, parada)
Place (lugar)	dental, coronal, labial, velar, glottal, alveolar, palatal, high, mid, low (dental, coronal, labial, velar, glotal, alveolar, palatal, alto, medio, bajo)
front-back (delante – atrás)	Front , back (delante, atrás)
Rounding (redondeo)	+ round, - round (redondeada, no redondeada)

"Voicing" describe la vibración de las cuerdas vocales, "manner" la manera en que se hace la articulación, "place" está referido al lugar donde se produce la articulación (la posición de la constricción en el tracto vocal durante la producción de las consonantes, o la altura de la lengua durante la producción de las vocales), "front-back" indica la posición de la lengua y "rounding" se refiere a la forma que toman los labios durante la producción del habla.

Cada una de estas clases abstractas se define como variable cualitativa que en dependencia del segmento de voz a analizar toma un valor de todos los posibles a tomar. Los valores de dichas variables cualitativas tienen en la mayoría de los casos una explicación sencilla de la articulación. Cada variable cualitativa debe incluir además el valor "silencio" y el valor "nulo" los cuales precisan que esa variable cualitativa no tiene importancia (no aporta información) en el segmento de voz que se analice.

El fenómeno de co-articulación es la modificación de un sonido fonético debido a la anticipación o la preservación de sonidos fonéticos adyacentes. Estas modificaciones se deben al mecanismo de producción del habla: Los sonidos que los oyentes identifican como segmentos de habla se producen en forma concatenada debido a la superposición de los gestos articulatorios que se coordinan en paralelo. El tiempo en que se producen estos gestos, sin embargo, no es simultáneo pero si altamente imbricado, como se puede apreciar en la figura 5, que muestra el tiempo de articulación relativo del velo del paladar, del cuerpo y de la punta de la lengua, de los labios, y de la glotis, durante la producción de la palabra "pan" en el idioma inglés. Esto es una representación abstracta de los movimientos reales del articulador, como si hubiesen sido determinados por estudios de rayos X [61] . Los rectángulos representan la extensión temporal

de los movimientos de los articuladores listados en el eje vertical. Se nota que los gestos producidos por los diferentes articuladores que provocan los segmentos fonéticos en el eje horizontal no siguen modelos temporales idénticos.

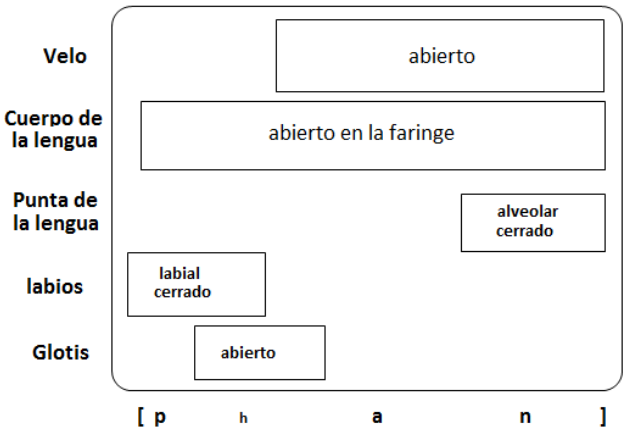

Fig. 5. Tiempos de producción relativos a los gestos articulatorios que tienen lugar en la producción de la palabra "pan" en el idioma Inglés

Debido a la inercia propia de los articuladores, las combinaciones articulatorias no cambian rápidamente de un segmento de habla para el siguiente. Más bien, los gestos articulatorios evolucionan relativamente lentos en el tiempo y típicamente cubren períodos de tiempo conteniendo varios segmentos de habla. Las propiedades espectrales de estos segmentos son consecuentemente afectadas por la manera en que el gesto cambia las propiedades de resonancia del tracto vocal. Estas modificaciones espectrales son perceptibles causando un cambio en la identidad del segmento, pero poco distintivos, por lo que pueden tener efecto sobre el modelo estadístico del sonido en cuestión durante la tarea de identificación. Todas las instancias de co-articulación influencian la representación acústica de los sonidos fonéticos y pueden obscurecer sus modelos acústicos.

Los enfoques tradicionales que se basan en los modelos acústicos de los sonidos fonéticos ignoran la fuente de información real que constituye la co-articulación y las ventajas potenciales que se ganan de una descripción directa de esta fuente. Estudios de los fenómenos articulatorios [54, 61] han mostrado que la mayoría de los fenómenos de co-articulación pueden ser ubicados en una reorganización temporal y/o espacial de los gestos articulatorios. Dichos gestos se pueden superponer por largos períodos de tiempo debido un incremento en la velocidad al hablar, o pueden tener una magnitud más pequeña si el habla es lenta. Si cupiera construir una representación fidedigna del proceso articulatorio, los fenómenos de co-articulación podrían ser modelados simplemente en términos de estas manipulaciones básicas de gestos articulatorios. En el dominio espectral o cepstral, en contraste, estas modificaciones de los gestos pueden generar patrones complicados que son difíciles de interpretar o de modelar.

Las características articulatorias son descritas tanto para la señal acústica como para unidades lingüísticas en un nivel más alto, como los fonemas y sílabas. Por consiguiente proveen un lenguaje de descripción más adecuado para las variantes de pronunciación, permitiendo que las palabras en el léxico del reconocimiento no estén representadas en términos de las secuencias fonéticas rígidas sino en términos de las secuencias paralelas de características articulatorias que están holgadamente sincronizadas.

Por supuesto, es razonable suponer que los diferentes parámetros articulatorios poseen diferentes grados de robustez, y no se deterioren bajo condiciones acústicas adversas. Las variaciones del parámetro "*voicing*", por ejemplo, pueden ser detectadas con una mediana robustez a través de variadas condiciones acústicas [62]. La detección de la característica "*place*", en contraste, es probablemente menos robusta ya que requiere recobrar el punto de la constricción articulatoria en el tracto vocal mediado por la señal acústica. Los cambios acústicos inducidos por las diferentes posiciones de constricción, sin embargo, dependen en gran medida de las características del tracto vocal de los oradores. Estas condiciones hacen factible la posibilidad de explotar estas características teniendo en cuenta diferentes grados de selectividad entre ellas.

A pesar de las ventajas que traería la inclusión de las representaciones articulatorias descritas anteriormente, se debe especificar que: en primer lugar, es difícil encontrar una fuente fidedigna a partir de la cual extraer los parámetros articulatorios de la señal acústica; En segundo lugar, el uso de información articulatoria requiere procesamiento adicional, lo que constituye una traba para la integración de este acercamiento en aplicaciones de gran escala.

Otra razón por la que las representaciones articulatorias no han sido usadas extensamente en sistemas de reconocimiento del lenguaje hablado es el costo adicional asociado con ellos. Si los parámetros articulatorios son extraídos estadísticamente de la representación acústica, se requeriría una fase adicional de elaboración. Calcular una transformación inversa acústica – articulatoria haciendo mapas es análogamente caro, puesto que esa tarea usualmente requiere una búsqueda del espacio de las características articulatorias. Sin embargo, este coste adicional puede ser aceptable si simplifica el procesamiento a nivel más alto, como la evaluación de modelos acústicos o la búsqueda léxica durante la decodificación. Además, una arquitectura modular, paralela de reconocimiento puede ser capaz de manejar los requisitos adicionales de procesamiento en el tiempo real.

6 Rasgos obtenidos de la señal del habla. Principales métodos para su extracción automática

Existen diferentes formas de extraer automáticamente las características de la señal del habla digitalizada. 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 acústicos que mostraron resultaron ser provechosos en la identificación de locutores con el objetivo de ver si eran representativos de 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 de LID han estado orientados a la fusión (concatenación) de varios tipos de rasgos, creando vectores de características cada vez más complejos. A continuación se explica brevemente los rasgos que más se han utilizado y las tendencias actuales de su utilización.

6.1 Rasgos acústicos

6.1.1 Rasgos lineales. Coeficientes de predicción lineal (LPC)

El método de predicción lineal o LPC (Linear Prediction Coeficient), también conocido como método de modelación autorregresiva, fue propuesto por Atal [63] para la codificación de la voz y generalizado por Markel [64] para el análisis de la voz. Consiste en simular la estructura del tracto vocal, pasar una señal periódica a través de este y obtener los coeficientes que se ajusten a la característica espectral de la señal resultante. Con una cantidad determinada de coeficientes se logra una buena aproximación del espectro de una gran cantidad de sonidos. Este método se utilizó por primera vez en el reconocimiento del locutor para obtener los rasgos acústicos [65] y brindan una información combinada sobre el contenido de frecuencia, ancho de banda de los formantes y la onda glotal.

Este método se fundamenta en la suposición de que la forma del tracto vocal se aproxima a la de un tubo acústico compuesto por varias secciones cilíndricas de diferentes diámetros (fig. 6) y se extiende desde la glotis hasta los labios sin ramificaciones. La onda de sonido se propaga y refleja en toda la extensión del tubo acústico provocando resonancias, su comportamiento se describe con la función transferencial de un filtro todo-polo.

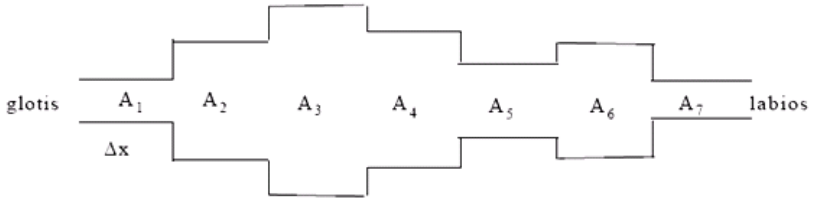

Fig. 6. Modelo del tracto vocal

La necesidad de encontrar otras variantes de este modelo está relacionada con la simplificación del procesamiento computacional. Mientras menos correlacionados estén los coeficientes obtenidos como resultado del modelo, mayor será la información que ellos contienen. Pero si los coeficientes están altamente correlacionados, se reflejara redundancia en los datos y esto provoca que baje la eficiencia del procesamiento. Es por eso que se han aplicado diferentes transformaciones sobre los rasgos LPC que dan lugar a varias familias de coeficientes menos correlacionados. Por ejemplo los coeficientes de reflexión, los coeficientes de razón logarítmica de áreas [66], los coeficientes de reflexión Arcsin y los coeficientes de pares de líneas espectrales [6] [67].

Las ventajas fundamentales de este modelo radican en la precisión de las estimaciones obtenidas y su rapidez de cálculo [68] . Sin embargo, para la definición del modelo LPC se han hecho una serie de asunciones poco rigurosas que redundan en inconvenientes para su aplicación. Según Atal [65] el tracto vocal no está compuesto solo por cilindros y la cavidad nasal constituye un pasaje adicional. Además la señal sonora que conforma la voz no tiene una estructura espectral plana. Por otro lado, son poco útiles en ambientes reales, pues presentan problemas de robustez ante la degradación del habla producto del ruido y ante cambios en el canal [69]. Además de lo expuesto, de forma general, los resultados experimentales obtenidos con estas familias de coeficientes son bastante similares y no muy efectivos en la identificación de idiomas.

6.1.2 Rasgos cepstrales

Los rasgos cepstrales son una mejor aproximación matemática del habla que el modelo LPC, aunque es importante señalar que no contienen información sobre la variación en el tiempo de las características propias de la señal de voz.

El cepstrum de una señal se define como la transformada inversa de Fourier del logaritmo del espectro de potencia de dicha señal [6]. A partir de este hecho queda definido un nuevo dominio, diferente al dominio del tiempo y al de la frecuencia, el dominio del cepstrum [70].

Los rasgos 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 los rasgos LPCC son los rasgos cepstrales más populares.

Los MFCC son utilizados para la representación del habla basándose en la percepción auditiva humana. Para su obtención se parte de un filtrado en bandas del espectro de potencia, teniendo en consideración que las bandas de frecuencia del filtro se espacian logarítmicamente, según la escala de Mel (escala logarítmica desarrollada para representar mejor el patrón de percepción del oído humano). Para obtener los coeficientes MFCC se le aplica al espectro de la señal de voz un banco de M filtros triangulares espaciados según la escala Mel que generalmente se extiende sobre toda la banda de frecuencias de la señal de voz como muestra la figura 7, aunque según el caso se valora la posibilidad de atenuar algunas zonas del espectro que no son útiles para el procesamiento. Como resultado del filtrado se obtienen un espectro de potencia distorsionado en escala Mel. Luego se le calcula el logaritmo para llevarlo al dominio cepstral. Finalmente se calculan los coeficientes cepstrales utilizando la transformada discreta del coseno (DCT). La cantidad de coeficientes cepstrales se corresponde con el número de filtros (M) del banco en escala Mel. Estos pueden variar entre 24 y 40 según la implementación [45].

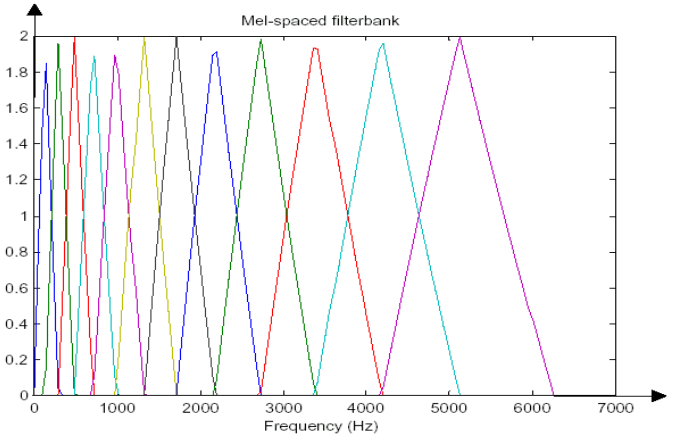

Fig. 7. Banco de filtros espaciado según la escala Mel

Para el reconocimiento del idioma usualmente se escogen 12 coeficientes cepstrales. Davis y Mermelstein [71] fueron los primeros en aplicar los MFCC para el reconocimiento del habla con exitosos resultados. La misma representación fue adoptada por los investigadores para el reconocimiento del locutor, demostrando ser altamente exitosa en esta área [72-74].

6.1.3 Rasgos dinámicos ceptrales. Rasgos delta (Δ) y delta-delta (Δ Δ)

Los órganos articulatorios cambian constantemente de forma durante el proceso del habla. Estos movimientos se reflejan en el espectro de la señal de voz como cambios en las frecuencias y anchos de banda de los formantes. A los rasgos que contienen la información de la señal teniendo en cuenta su variación en el tiempo se les denomina rasgos dinámicos. 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 [75]. Comúnmente se estiman también las derivadas en el tiempo de los rasgos delta, conocidas como rasgos delta-delta (Δ Δ), y se anexan al vector de rasgos.

Como fuente para calcular los rasgos dinámicos se puede utilizar cualquier rasgo espectral, especialmente los rasgos cepstrales y sus variantes [67]. Aplicar los métodos de extracción de rasgos delta y delta-delta, tomando como base los rasgos cepstrales (MFCC), completa la información que necesita un ASR para escalar en eficiencia. La desventaja es que ellos provocan que el vector de rasgos aumente su dimensión de forma considerable, lo que resta a la hora de valorar el costo del procesamiento matemático y computacional.

La cantidad de tramas necesarias para obtener el mejor estimado de la dinámica cepstral está dada según el orden del rasgo cepstral, pues cada rasgo tiene su propia dinámica temporal. Pocas tramas dan como resultado un estimado ruidoso. A medida que se aumenta el estimado se suaviza, lo que tampoco es conveniente. Por lo tanto hay que llegar a un compromiso [45].

6.1.4 Rasgos Cepstrales Delta Desplazados (SDC)

Los rasgos SDC o Shifted Delta Cepstral en inglés, se utilizan frecuentemente en la identificación de idioma y dialecto con 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) [76, 77]:

N: dimensión de los vectores de rasgos cepstrales (MFCC).
d: cantidad de tramas a escoger para calcular el rasgo delta (Δ).
P: número de tramas 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, para cada valor de t, se toman k vectores de N rasgos delta (Δ), 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 Δc (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 dimensión de los vectores de rasgos fuente, k es el número de rasgos delta (Δ) que se toman para formar el SDC y t el instante de tiempo en que se toman las muestras. Los valores

de d y P intervienen en el conjunto de componentes del rasgo fuente, e influyen en la formación de cada 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 deltas (Δ) pero incluyendo el desplazamiento en tiempo iP entre bloques [78]:

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

Es importante destacar que con esta técnica se tienen 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 la dinámica de la señal, reflejando las transiciones entre fonemas e incluso sílabas, lo que conduce a una mejora en la efectividad de la verificación del lenguaje.

Los rasgos cepstrales portan información sobre la estructura instantánea del tracto vocal y por ende de los formantes del habla, entonces los rasgos SDC reflejan la dinámica de dichos rasgos, o sea el movimiento y posición de las articulaciones vocales y nasales que controlan la dinámica del movimiento del tracto vocal. Si el intervalo de tiempo en el que se evalúa la señal es suficientemente largo para incluir las transiciones espectrales entre fonemas y sílabas, se puede decir que los rasgos SDC reflejan la dinámica temporal de las articulaciones durante el habla, reflejando un comportamiento pseudo-prosódico, lo que puede justificar los buenos resultados obtenidos en reconocimiento de idioma y dialecto [45].

6.2 Rasgos prosódicos

No solo la fisiología de los órganos que producen la voz influye en la forma de hablar de una persona. Los hábitos influyen mucho en la manera particular de expresarse de cada cual, y son características adquiridas en un período de tiempo determinado que están muy influenciados por el marco social y por las particularidades del idioma que se aprende en la infancia. Esas particularidades de los hablantes y por supuesto del idioma es lo que recoge, en cierto modo, la prosodia, lo cual desempeña un papel crucial en la percepción humana del habla. En los últimos años, cada vez más los investigadores del área de reconocimiento de los idiomas muestran interés en las características prosódicas ya que ellas encierran lo que los especialistas lingüistas llaman información suprasegmental [3].

Generalmente, la prosodia quiere decir "la estructura que organiza el sonido" [79]. La información contenida en los rasgos prosódicos es en parte diferente a la información contenida en las características cepstrales. Los rasgos prosódicos representan la entonación, el acento, las pausas y el ritmo, lo cual determinan el estilo y la cadencia del habla. Desde el punto de vista acústico, existen tres características físicas de la señal de voz que son de naturaleza prosódica y se refieren al armónico fundamental de la señal. Estas son: la frecuencia fundamental, la intensidad y la duración. Los rasgos prosódicos aparecen a la hora de medir estas características constituyendo atributos perceptuales de las características físicas antes mencionadas [80], que se les conoce como tono (por su nombre en inglés: pitch), energía y tempo.

Recientemente se han desarrollado varias técnicas para capturar la información suprasegmental existente en las señales de habla, para utilizar esta información en la identificación de idiomas.

Estos métodos han incluido el cálculo del contorno del tono fundamental , en aprendizaje no supervisado a través de GMM, con la fusión con otros rasgos [81]; y en aprendizaje no supervisado través de HMM [82]. Cummins et al [83] basan el procesamiento acústico solamente en el uso de la *frecuencia fundamental*. Samouelian [84] amplia el conjunto de características suprasegmentales al usar frecuencias cepstrales secundarias, calculando los cambios o deltas en ventanas adyacentes de 12 rasgos MFCC, ya que los cambios temporales en el espectro juegan un papel muy importante en la percepción humana, una de las maneras de capturar esta información es el uso de los deltas Δ en los coeficientes cepstrales. Con ellos es posible describir el cambio de cada coeficiente en el tiempo, capturando así los cambios entre fonemas, sílabas y quizás hasta palabras, buscando de esta forma capturar parte de la información suprasegmental.

Ana Reyes [3] 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 se plantea capturar los cambios temporales en el espectro de la señal de voz por medio del uso de los coeficientes cepstrales MFCC, con 16 coeficientes, capturando además de la frecuencia fundamental F0, frecuencias secundarias por medio 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, que con muestras pequeñas de señal de voz brindo mejores resultados en la identificación del lenguaje, al contrario de lo que se esperaba puesto que los deltas tienden a estabilizarse para muestras de señal de voz más grandes. Esto ocurre probablemente por la mayor aparición de pausas y silencios, cuando se toman muestras pequeñas.

El segundo método propone obtener una nueva caracterización orientada a las bajas frecuencias, partiendo de que la frecuencia fundamental es la más baja de todas y es el parámetro más utilizado para representar la prosodia. Por lo tanto se asume que en las frecuencias bajas hay información relevante para la identificación del idioma y 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 es que la ventana de la transformada de Fourier de tiempo corto (STFT) es de duración fija, mientras que la de wavelet es de duración variable, lo que permite establecer el grado de resolución tanto de tiempo como de frecuencia. Para tener una buena resolución de las altas frecuencias se necesitan ventanas pequeñas de tiempo y para una buena resolución de las bajas frecuencias se necesitan ventanas grandes de tiempo, lo que resulta importante, ya que se desea capturar la información suprasegmental del habla.

Otra diferencia a considerar es que el método wavelet es sensible al tamaño de muestra de la señal de voz, pues se necesitan muestras de voz de alrededor de 50 segundos para obtener 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 [85, 86], pero hasta ahora no se había reportado ningún trabajo aprovechando el uso de wavelet en la identificación de idiomas. En general la obtención de los rasgos prosódicos del habla requiere gran cantidad de muestras de voz y por eso consumen mucho tiempo de procesamiento.

6.3 Características Modificadas de Retraso del Grupo (MODGDF)

Las investigaciones en los últimos años han estado orientadas a trabajar con las características de magnitud del espectro, 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. 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 desde el punto de vista espectral. De acuerdo con esto el lenguaje hablado puede ser entonces caracterizado por información ya sea de magnitud o de fase [87].

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 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 [87], y su función de retraso del grupo se parecen.

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: [88].

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

Donde los subíndices R de e I denotan las partes reales e imaginarias de la transformada de Fourier. X(w) y Y(w) son las transformadas de x(n) y y(n), respectivamente.

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

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

Donde N_f es el orden de la DFT y $r_x(k)$ es la función de retraso de grupo. Específicamente se usa la segunda forma de la transformada discreta del coseno, aprovechando que tiene propiedades asintóticas (Karhunen –Loeve, KLT). La DCT se utiliza para eliminar la correlación lineal, lo cual permite el uso de covarianzas diagonales a la hora de modelar el vector c(n), que es quien constituye 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 procesamiento del habla [88], así como en la identificación de locutores [89], y el reconocimiento de sílabas [90].

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 [91], como espontáneas [92]. Se demostró que los rasgos MODGDF mejoran en un 3% el rendimiento de los sistemas LID, con respecto a los rasgos MFCC, y que su concatenación es aun mejor. Mostraron además tener un mejor funcionamiento en términos de separación de clases cuando se utilizan métricas como Bhattacharya en el proceso de clasificación [93].

6.4 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 de las señales de habla, como la información de fase. Sin embargo, hay que tener cuidado porque los rasgos 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.

Otra motivación a la búsqueda de nuevas características basadas en la fase fue el análisis del proceso de extracción de los rasgos MFCC, donde se notó que sólo se usa la magnitud del espectro para producir los coeficientes, dejando simplemente descartada la fase 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 sugirieren que la información de fase tenga una contribución significativa para la inteligencia audible humana [94]. Como alternativa a situación se propone el uso de los componentes de Modulación de Frecuencia (FM) de una señal de habla en un modelo AM-FM [95].

En esta nueva alternativa las resonancias del tracto vocal son modeladas como señales de modulación de amplitud y frecuencia en un modelo AM-FM, dicho modelo, sugiere la descomposición de señales de voz en una serie de señales moduladas por la frecuencia instantánea y la amplitud. Considerando que las distribuciones de frecuencia y tiempo de las señales del habla contienen información acústica en la parte no lineal del espectro del habla; y que los resultados en ASR [96] indican que la información acústica no puede ser modelada solo 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, basadas ya sea el la modulación de frecuencia FM o en la modulación de amplitud AM, para proveer información acústica adicional.

Para aislar el componente de frecuencia modulada en cada resonancia, de los demás, se aplica un filtro de paso de banda a cada componente de resonancia. Existen varios métodos, incluyendo al popular algoritmo DESA [97] y los métodos basados en las transformadas de Hilbert [98].

Sin embargo, las estimaciones de frecuencia modulada producidas por estas técnicas no son suficientemente coherentes debido a picos que salen a relucir ocasionalmente [97, 99]. Consecuentemente la clasificación no puede ser lograda con suficiente exactitud, usando como características, estas estimaciones de frecuencia modulada.

Para producir características más consistentes, basadas en modulación de frecuencia, fue propuesta una técnica alternativa en [99] en la que la señal de cada banda es modelada por un resonador todo-polo de segundo orden, 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. Los componentes de frecuencia modulada de todas las bandas son entonces concatenados para formar un vector de características de FM [100].

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.

6.5 Rasgos Articulatorios

Son varias la investigaciones [55, 101, 102] que han demostrado que las características basadas lingüísticamente en la relación directa del proceso de articulación humano pueden captar mucho mejor las características del discurso del habla.

Estos rasgos articulatorios (AF) pueden reemplazar o complementar las características acústicas en el procesamiento de señales de voz y pueden ser definidos como descripciones abstractas de importantes propiedades del tracto vocal y de los movimientos articulatorios que tienen lugar durante la producción del habla.

Recientemente, ha habido un interés renovado en aplicar este tipo de información como características suplementarias o alternativas para las tareas de procesamiento del habla [55, 56, 101, 103-106] y se han tomado acuerdos generales sobre las propiedades articulatorias de las unidades fonéticas, desarrollándose diversas formas para representar o codificar estas propiedades, tal que puedan ser extraídas y modeladas. El problema es a menudo, el hecho de que muy pocas bases de datos miden directamente el movimiento y la posición de los articuladores humanos, en lugar de eso, sólo se conocen las posiciones teóricas del estado estable de los fonemas. Zhang y Edmondson [107] propusieron características articulatorias de valor continuo las que regularmente son discretizadas, porque aunque un valor continuo tiene mejor "resolución", este es difícil de estimar a partir de bases de datos con que se cuenta. A continuación se describen los principales métodos de extracción y modelación que se han usado bajo el enfoque de las AF.

6.5.1 Clasificación por descomposición en términos de rasgos articulatorios

La figura 8 muestra un esquema representativo de la aproximación a las características articulatorias a partir de un modelado acústico. La idea básica de este enfoque es usar una representación de la señal del habla en un estado intermedio entre la señal acústica pre-procesada y el nivel de estimación de probabilidad de la unidad de sub-palabras, que refleja una relación con el proceso articulatorio subyacente en la señal. Las probabilidades de las características articulatorias son extraídas a partir de la señal acústica pre-procesada por un conjunto de clasificadores estadísticos independientes trabajando en paralelo. En un segundo paso, esta representación de las características articulatorias se mapea como puntuaciones de unidades de sub-palabra de más alto nivel, tales como sonidos fonéticos o sílabas.

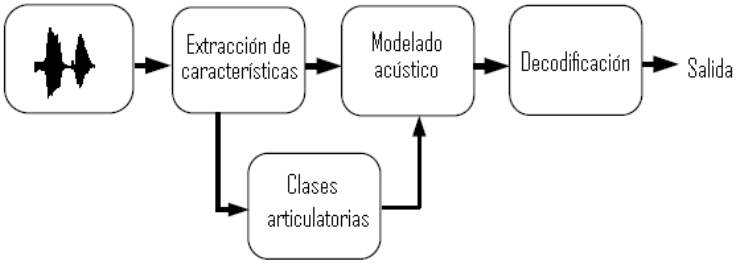

Fig. 8. Sistema de reconocimiento que incluye la representación articulatoria

La elección particular de las características articulatorias y la disposición de sus correspondientes clasificadores está basada en la estructura de producción del habla en los humanos. La producción del habla en los humanos implica la interacción de varios componentes o dimensiones articulatorias que son parcialmente independientes unas de otras. Por esta razón

se proponen clasificadores separados para estas dimensiones articulatorias, tal como se muestra en el esquema de la figura 9. Se puede distinguir bien entre las maneras de articulación, ej. la forma de la contracción hecha por el articulador en el tracto vocal, y el lugar de la articulación (la localización de la contracción). La primera de estas dimensiones "voicing", la cual describe el estado de las glotis y la actividad de las cuerdas vocales, altamente dependiente de la actividad articulatoria en el tracto oro-nasal. La quinta dimensión articulatoria, "rounding" de los labios, es altamente independiente de la mayoría del cuerpo de la lengua o de los movimientos de la punta de la lengua y pueden afectar grandes tramos de la señal de habla. Finalmente la posición relativa de la lengua "front-back" de los dientes es otra propiedad articulatoria que también muestra un comportamiento independiente temporalmente. Algunas de estas dimensiones articulatorias no son enteramente independientes: aunque las mayoría de las formas de contracción, por ejemplo, pueden producirse en las mayoría de los puntos en el tracto vocal, hay ciertos lugares de articulación que son compatibles con ciertas maneras de articulación, ej: por ejemplo no hay consonantes glotales que tengan una forma lateral de constricción. Estas interdependencias se resuelven gracias a la distribución paralela de los clasificadores de características. Los clasificadores de alto nivel cuyo funcionamiento se basa en el mapeo de la representación de las características articulatorias a lo largo de las unidades de sub-palabras deben ser capaces de aprender restricciones de las co-ocurrencias de ciertas características articulatorias a partir de los datos.

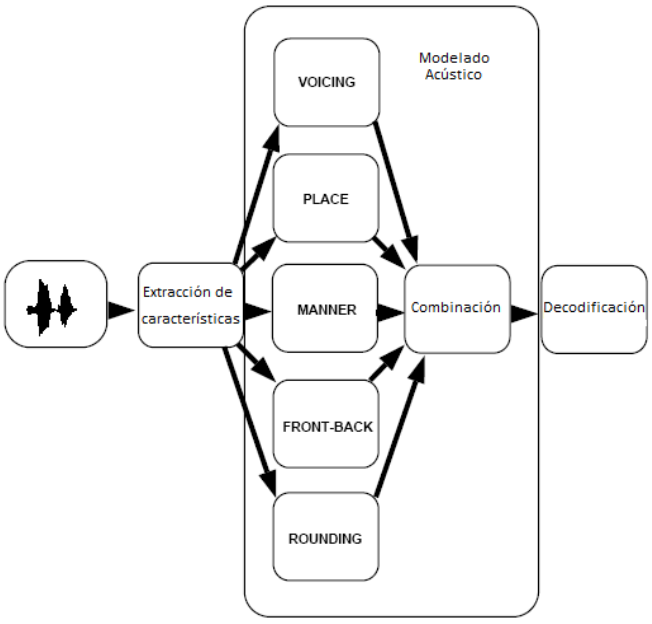

Fig. 9. Enfoque de modelación acústica a partir de características articulatorias

Ahora bien, la tarea de clasificar un vector de observación acústica x como perteneciente a una de las diversas clases fonéticas es muy complejo debido a la variabilidad en la señal de habla. Para *N* fonemas, esta clasificación incluye la estimación de *N* probabilidades de un fonema dada una observación acústica x, *P(fonema|x)* (o dependiendo del clasificador , las *N* vecindades de la observación x dado un fonema, *P(x|fonema)*). Asumiendo que las unidades de sub-palabras en cuestión son fonemas independientes del contexto, *N* generalmente está en el rango entre 30 y 60. El esquema descrito anteriormente en la figura 9 muestra una estructura de clasificadores en cascada, cada clasificador articulatorio de alto nivel solo necesita distinguir entre un número pequeño de clases resultantes. Generalmente, el número de clases requerido

describe la dimensión articulatoria en un rango de 2 a 3 (para “voicing”) y hasta aproximadamente 10 (para “place”). Los datos de entrenamiento para los clasificadores articulatorios se generan convirtiendo las trascripciones fonéticas de entrenamiento en transcripciones de características articulatorias. Esto se puede hacer usando una tabla de conversión canónicamente definida. Como las características articulatorias pueden generalmente ocurrir en más de un fonema, los datos de entrenamiento para estas características pueden estar efectivamente compartidos por los fonemas, pero esto requiere una gran cantidad de material de entrenamiento para cada clasificador de características, que a menudo excede la cantidad de material de entrenamiento fonético en un orden de magnitud [55].

Es por eso que los diferentes aspectos de articulación exhiben diferentes grados de robustez y no se deterioran (en términos de su habilidad de estar reconocidos correctamente) bajo condiciones acústicas adversas. Esta estructura de clasificación basada en la descomposición de sonidos del habla en sus componentes articulatorias puede explotar esta propiedad aplicándose selectivamente diferentes estrategias de procesamiento a los diferentes sub-clasificadores, de forma independiente. Estas estrategias pueden implicar por ejemplo el uso de diferentes técnicas de adaptación al modelo. La característica “voicing”, por ejemplo, pueden ser detectada de forma medianamente robusta a través de una variada colección de condiciones acústicas [62]. La característica “place”, en contraste, tienden a ser menos robusta puesto que depende más de las características del tracto vocal de los oradores. Así se podría sacar provecho de un método de adaptación al modelo que sería aplicado solo al clasificador de “place”.

Los clasificadores articulatorios pueden diferir también en el tipo de clasificador, en la complejidad (el número de parámetros libres) y en la inicialización o en los métodos de entrenamiento que pueden afinarse para la tarea específica que necesitan realizar. Además de usar estrategias selectivas de procesamiento y adaptación en la primera etapa de clasificación, las contribuciones de los sub-clasificadores para la tarea global de clasificación pueden ser balanceados en el módulo de combinación, en dependencia del contexto. El módulo de combinación, por ejemplo, puede destinar valores de confianza como base para asignarle pesos a las salidas de los sub- clasificadores. Por todas estas razones, un acercamiento acústico de la modelación que se base en la clasificación por descomposición de rasgos AF tiene probabilidad de resultar ser más robusto ante condiciones acústicas adversas. [108]. A continuación se describen las principales aproximaciones a la extracción de AF.

6.5.2 Mapeo articulatorio – acústico y mapeo inverso

Para modelar la relación existente entre los parámetros articulatorios y los parámetros acústicos de la señal se pueden construir mapas en ambas direcciones entre estos dos conjuntos de parámetros. Uno es el mapeo articulatorio-acústico, es decir, la determinación del espectro acústico a partir de los parámetros articulatorios, y otro es el mapeo inverso acústico-articulatorio, es decir, la determinación de los parámetros articulatorios a partir de la configuración de la señal de voz. Hay una diferencia fundamental entre estos dos mapeos. En general, se puede obtener un mismo sonido usando diferentes configuraciones articulatorias, o sea diferentes combinaciones de parámetros articulatorios podrían corresponder al mismo espectro acústico (problema uno a muchos).

El mapeo articulatorio-acústico es ampliamente usado en aplicaciones de codificación, síntesis y modificación de habla. El mapeo inverso acústico-articulatorio es también usado por aplicaciones como visualización y codificación de habla con baja razón de bits.

Los enfoques anteriores al mapeo articulatorio-acústico y al mapeo inverso, empleaban parámetros articulatorios definidos por un modelo matemático, [109, 110] en el que la señal de

voz se generaba a partir del mismo. Estos planteamientos tienen la limitación fundamental de que el mecanismo de producción de la voz es demasiado complejo para ser matemáticamente modelado sin una cierta aproximación.

Uno de los primeros sistemas que hace uso de parámetros articulatorios fue el reportado por Schmidbauer [111] quien desarrolló un reconocedor de habla para alemán usando 19 características articulatorias que describían la manera y el lugar de la articulación. Estas fueron detectadas del pre-procesamiento de la señal de habla por medio de un clasificador de Bayes, las probabilidades posteriores de las características fueron concatenadas en un vector de rasgos articulatorios, que fue entonces utilizado como entrada de datos a los modelos fonéticos (HMM).

Dalsgaard y colegas. [112, 113] proponen tres valores de parámetros articulatorios para el etiquetado multilingüe. Una red neural auto-organizada (SONN) detecta las características y su salida se utiliza como entrada para los modelos de fonema multivariados creados con un clasificador GMM, y estos a su vez, sirven para la alineación automática de etiquetas por un algoritmo Viterbi.

Probablemente el sistema de parámetros articulatorios más elaborado ha sido desarrollado por Deng y colegas [56, 114, 115]. Los autores usaron 18 características para describir cuatro dimensiones: sonorización, lugar de articulación, movimiento vertical del cuerpo de la lengua y movimiento horizontal del cuerpo de la lengua. Además, modelaron la señal del habla como una secuencia de vectores articulatorios intercalando vectores transicionales. Los vectores articulatorios estaban definidos por una combinación basada en reglas de los parámetros articulatorios y los transicionales estaban preparados para asumir cualquier valor fonéticamente verosímil entre el valor del vector previo y el siguiente. Los vectores se corresponden con los estados HMM estando combinados en un solo HMM ergódico cuyas transiciones y emisiones fueron entrenadas.

Kirchhoff en 1999 [55] entrenó un clasificador para cada propiedad articulatoria con el objetivo de aprender la correspondencia entre las características espectrales y los correspondientes estados de articulación. Para ello se entrenaron cinco perceptrones multicapa (MLP) cuyas salidas representaban las probabilidades a posteriori de las clases definidas para cada propiedad articulatoria. Posteriormente se adoptó una función softmax como la función de activación de salida de los MLP.

Las entradas para los MLP eran idénticas, mientras que su número de salidas varía en dependencia del número de clases definidas por cada parámetro articulatorio. Para garantizar que las estimaciones de los valores de los parámetros articulatorios fueran lo mas exactas posibles, para evaluar una trama t se toman como entrada a los MLP varias tramas de rasgos MFCC con índices de tramas consecutivas desde t –n/2 hasta t+n/2. En otras palabras, las AF en la trama t se obtienen a partir de n tramas consecutivas de rasgos MFCC centrados en la trama t. Antes de predecir las parámetros, los MFCC se normalizan a media cero y varianza unitaria con el objetivo de limitar las entradas del MLP dentro de un rango adecuado, de modo que la determinación de los pesos del MLP no esté dominada por las entradas de gran magnitud.

Los MLP pueden ser entrenados a partir de datos de voz con etiquetas fonéticas alineadas en el tiempo. Las alineaciones fonéticas se pueden obtener de la transcripción o la decodificación de Viterbi utilizando modelos de fonemas. Con las etiquetas de fonemas, y los tipos de articulación se deriva entonces la correspondencia entre los fonemas y sus estados de articulaciones. En la fase final el resultado de los MLP pasa a un clasificador de nivel superior para llegar a una clasificación final.[55]

Este mismo enfoque ha sido utilizado en muchas investigaciones posteriores para la extracción y modelación de los parámetros articulatorios, variando en cada caso los clasificadores usados para las características y el clasificador de alto nivel, aplicando medidas de

confianza [116] y aplicándolo a la identificación de idiomas, obteniéndose resultados bastante alentadores [59, 108, 117]. En estos casos para cada parámetro se construye un modelo acústico estadístico, análogo a los modelos acústicos fonéticos. La decodificación acústica se efectúa independientemente para cada grupo de características, produciendo flujos paralelos de característica fonéticas. Para cada flujo de características, se estima un modelo individual de n-gramas de características. Luego durante la prueba, la señal pre-procesada es pasada a través del banco de reconocedores de características, y por el modelo n-grama de característica, con los cuales se calculan las probabilidades de cada flujo de características dado el idioma. Las puntuaciones específicas resultantes se combinan para proveer la puntuación global del LID.

Supphanat y Julie [118] usaron este mismo acercamiento con el objetivo de identificar idiomas en el 2006. La figura 10 muestra un diagrama en bloques del sistema propuesto en su trabajo. Ellos se plantearon la obtención de modelos de parámetros articulatorios "AF" independientes del contexto, a partir de las trascripciones articulatorias usando HMM. Los modelos de características independientes del contexto son expandidos a modelos dependientes del contexto usando técnicas de apoyo. El lexicón de fonemas es también convertido en un lexicón AF según una tabla de características y se entrena el modelo del idioma usando un fondo de bigramas. El proceso de reconocimiento es igual al explicado en el trabajo de Kirchhoff [55], con la diferencia de que en este método se usa la información correspondiente a un solo parámetro articulatorio, tratándose de identificar para cada idioma cual de todos ellos es el que aporta mayor información. Como resultado de este trabajo se disminuyó el costo computacional del sistema (puesto que el acercamiento AF-HMM es muy sencillo, y se usan pocos parámetros) y se identificó que la característica articulatoria que mas información aporta a la identificación del idioma inglés y del Thai es el "type" referida fundamentalmente a las vocales. En este trabajo se presenta además una tabla de transcripciones fono-articulatorias muy completa y generalizada para varios idiomas [60].

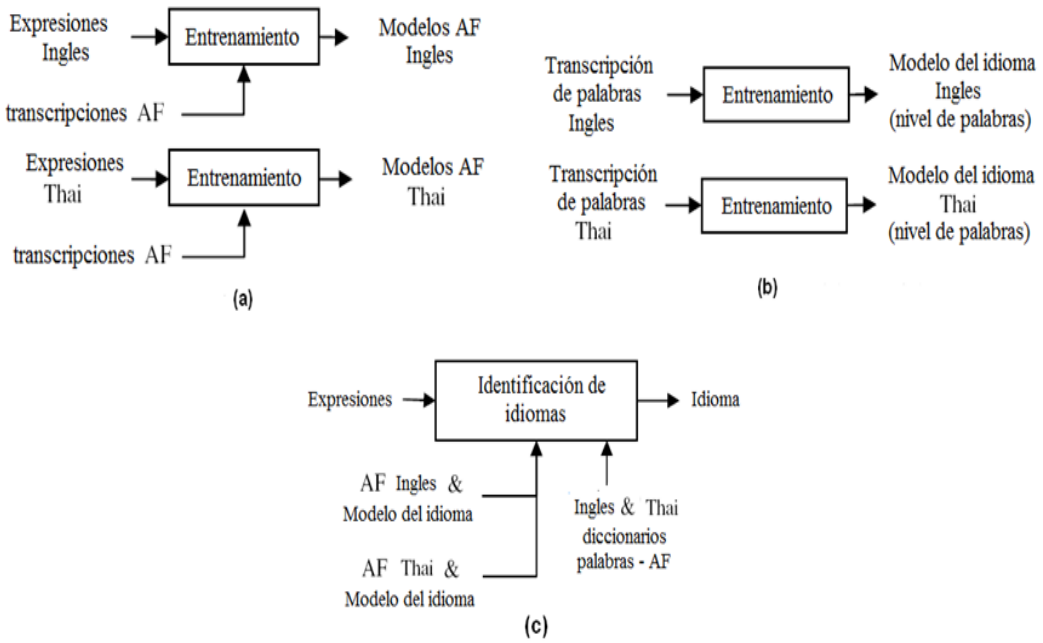

Fig. 10. Enfoque de un sistema de identificación de idiomas basado en AF. a) Entrenamiento de los modelos AF. b) entrenamiento del modelo del idioma. c) Proceso de identificación de idiomas

Este mismo grupo de trabajo se planteo la extracción de la característica articulatoria "manner" usando un clasificador SVM con un kernel lineal (SVMLight [119]) En contraste al acercamiento HMM, se le impuso al SVM una descripción estática, o sea se entrena sólo con los parámetros que describen la trama en cuestión y cuatro tramas adyacentes que son usadas como fuente informativa para la predicción. En este trabajo [120], se convierten directamente transcripciones de fonemas en transcripciones de referencias AF. En el experimento se usaron 180 características del habla para el clasificador que fueron extraídas para cada trama en cuestión y dos tramas adyacentes antes y después es decir, 5 tramas en total, de cada trama se extrajeron los valores de energía global, entropía global, 12 MFCC con $\Delta 1$ y $\Delta 2$. Las etiquetas para cada una de las instancias fueron obtenidas a partir de las transcripciones de los fonemas del habla. Como resultado se obtuvo que el SVM solo superó al tradicional HMM en la tarea de extracción de la característica "manner".

Para la comparación se usó un método de evaluación de exactitud global en términos de la tasa de errores a nivel de trama (FER), ya que, actualmente las características articulatorias son comúnmente utilizadas como una representación alternativa del habla y su representación es una secuencia de vectores numéricos donde cada vector numérico representa tramas del discurso en el tiempo. Por consiguiente, los sistemas de extracción de rasgos AF son usualmente evaluados a nivel de tramas. [120]

Es evidente que cada vez se hace más indispensable contar con una colección de parámetros articulatorios fiables para avanzar en esta área de investigación. Esta limitación va siendo superada por el desarrollo actual de dispositivos de grabación que permiten grabar la señal del habla y los movimientos de varios articuladores simultáneamente. Las cantidades suficientemente grandes de datos paralelos acústico-articulatorios disponibles hacen posible la aplicación de un enfoque basado en corpus con asignaciones entre los parámetros de articulación y la acústica del habla. En lugar de representar matemáticamente el mecanismo de producción, la correspondencia entre la configuración de los parámetros de articulación y el espectro acústico puede ser estadísticamente extraída de las bases de datos paralelos acústico-articulatorios del habla.

Se han propuesto varios corpus basados en métodos de mapeo articulatorio-acústico. Un método es utilizar un libro de códigos. Shiga [121] propuso un método estadístico para estimar el espectro acústico a partir de espectros de armónicos en varias configuraciones de articulación similares. Kaburagi [122] comprobó que la precisión de la estimación del espectro acústico se puede mejorar mediante el uso de la información fonética, así como la información de articulación. Hiroya y Honda [123] ampliaron el enfoque del libro de códigos con modelos estadísticos de los parámetros acústicos y de articulación utilizando modelos ocultos de Markov (HMM) en un modelo de producción del habla. Este modelo representa la correspondencia entre los parámetros articulatorios y acústicos como una aplicación lineal en cada estado del HMM. Para representar asignaciones no lineales, Nakamura [124] empleó mezclas gaussianas como salida de funciones de densidad de probabilidad de los estados HMM dependientes de contexto. Kello [125] aplica una asignación no lineal con una red neuronal para mapas articulatorios - acústicos.

En los últimos años también se han propuesto varios métodos basados en corpus para el mapeo inverso acústico–articulatorio. Hogden [126] propuso un libro de códigos acústico-articulatorio. Suzuki [127] introdujo contrastes dinámicos de los parámetros acústicos y de articulación en el mapeo inverso para mejorar su rendimiento, en concreto, utiliza varias tramas acústicas como elementos de entrada, y una medida de la captura de discontinuidades articulatorias para seleccionar la secuencia óptima de vectores de salida. Richmond [128]

modeló el mapeo utilizando una red neuronal sobre la base de la estimación de la densidad de la mezcla.

Se ha informado de que la representación múltiple de densidad de probabilidad de articulación es eficaz para la asignación de la inversión. Hiroya y Honda [123] proponen la asignación inversa articulatoria-acústica mediante un modelo de producción de habla basado en HMM, como se mencionó anteriormente. En este método, no sólo se utilizan las características dinámicas de los parámetros acústicos y articulatorios, sino que se utiliza también la información fonética que se requiere para la formación de los HMM como limitaciones para abordar el problema de uno-a-muchos.

Toda [129] construyó asignaciones estadísticas entre los movimientos articulatorios y el espectro acústico, sin el uso de información fonética. La aplicación de este acercamiento proveería varias ventajas como por ejemplo permitiría la modificación y decodificación del habla sin que influya el idioma. Inicialmente Toda aplicó un algoritmo de asignación que utiliza el modelo GMM propuesto por Stylianou [130]. Este algoritmo determina el parámetro articulatorio a partir de los parámetros acústicos, trama por trama, utilizando como criterio el promedio del mínimo error cuadrado (MMSE). Sin embargo, aunque funciona razonablemente bien, el mapeo basado en MMSE no es apropiado para el trabajo con múltiples distribuciones de la densidad de probabilidad, ya que hace caso omiso de las covarianzas de las distribuciones, aun cuando son diferentes unas de otras. Por otra parte, el proceso de mapeo trama por trama trae consigo la determinación de trayectorias inapropiadas de los parámetros.

Como solución a estos problemas Toda [129] propuso un mapeo basado en GMM, pero usando la estimación de máxima verosimilitud (MLE) y considerando además las características dinámicas. El mapeo basado en MLE permite la determinación de un parámetro de trayectoria con propiedades estáticas y dinámicas adecuadas. Como conclusión el uso del mapeo basado en MLE resultó una mejora significativa en la precisión del mapeo, tanto en la conversión articulación-acústica como en la asignación inversa. Para los experimentos usaron la base de datos de habla acústica–articulatoria (MOCHA) [131].

6.5.3 Modelos de Pronunciación Condicional basados en características articulatorias

Las técnicas de modelación de la pronunciación son ampliamente utilizadas en los sistemas de reconocimiento del habla. La solución a esta tarea se logra generalmente mediante la clasificación de la señal de voz en pequeñas unidades de sonido (o sub-unidades de palabra), y luego combinarlas en palabras, y, finalmente, frases y expresiones, la figura 11 muestra este proceso de forma simplificada. El “pegamento” que une las palabras a sus unidades de sonido correspondiente es el modelo de pronunciación.

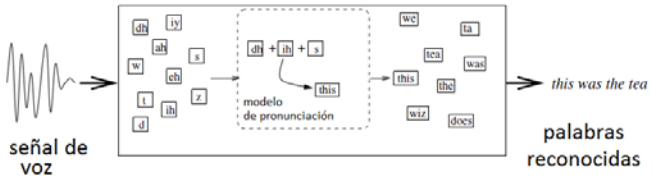

Fig. 11. Modelo de pronunciación en un sistema ASR

El modelo de pronunciación está profundamente arraigado en la ciencia de la lingüística, dos campos investigan los detalles de la producción de la pronunciación: la fonética, que estudia la

gama de sonidos vocales durante la generación de lenguaje hablado, y la fonología, que modela la variación fonética [132].

Los modelos de pronunciación determinan como las unidades de sub-palabras pueden combinarse para formar palabras. El sistema de reconocimiento del habla utiliza el concepto de fonema para representar las unidades de sub-palabras, por lo que el trabajo del modelo de pronunciación es determinar las secuencias de fonemas que constituyen palabras válidas.[133]

Esta técnica también ha sido empleada en tareas de reconocimiento de locutor. Entre todas las características de alto nivel evaluadas en [134], el mejor resultado se logró mediante un sistema que utiliza técnicas de modelado de pronunciación condicional "CPM"[135]. El CPM tiene como objetivo caracterizar el comportamiento de la pronunciación de un locutor mediante el cálculo de la correlación entre los fonemas previstos y la pronunciación producida por el locutor. Los fonemas previstos fueron obtenidos por un reconocedor limitado en léxico y las pronunciaciones reales fueron obtenidas por cinco reconocedores fonéticos sin gramática correspondientes a cinco idiomas diferentes [135]. Los comportamientos de pronunciación se codificaron como densidades de probabilidades discretas y estas fueron usadas en verificación de locutores de forma similar a los modelos GMM. Se comprobó que el modelado de pronunciación CPM es aplicable a la identificación de locutores porque estos tienen diferentes maneras de pronunciar un mismo fonema, y como resultado, un fonema realizado por distintos locutores puede ser reconocido como diferentes fonemas por los reconocedores fonéticos. Sin embargo, CPM no solo requiere datos de voz en varios idiomas para la formación de los modelos de fonemas, sino que también necesita largas expresiones para entrenamiento y la verificación del locutor.

Para evitar la necesidad de gran cantidad de datos multilingüe para entrenamiento, se propuso usar flujos de parámetros articulatorios para la construcción de modelos de pronunciación condicional CPM, aplicados a la identificación y verificación de locutores [57]. Aquí los parámetros articulatorios de cada locutor fueron modeladas por siete modelos dependientes del locutor, cada uno de los cuales modeló una clase de parámetro articulatorio, como una distribución condicional discreta. Para cada expresión, siete secuencias de clases articulatorias se obtuvieron de siete reconocedores basados en HMM, cada uno responsable de una propiedad de articulación.

En comparación con CPM basado en fonemas en [135], AF-CPM basado en parámetros articulatorios proporciona una unión más directa entre las variaciones de pronunciación y el proceso de producción del habla. Debido a que el proceso de producción del habla es una fuente de variación en los locutores, el AF-CPM basado en parámetros articulatorios resultó mejor que el CPM basado en fonemas, en términos de modelos de locutores. Además, las propiedades de articulación son las mismas independientemente del idioma y los datos de voz monolingües son suficientes para determinar sus valores.

En otro trabajo propuesto por el mismo equipo se reafirma [136] la utilidad de los parámetros articulatorios en la verificación del locutor. Esta vez, para cada expresión, se determinaron las probabilidades de las diferentes clases articulatorias a partir de cinco clasificadores basados en perceptrones multicapas "MLP" y se concatenaron para formar una secuencia de vectores de características articulatorias. Con este conjunto de vectores se procedió como tradicionalmente se evalúan los sistemas base GMM. En [136], se estimó la distribución discreta de cada modelo de locutor a partir exclusivamente de los datos de entrenamiento de los locutores correspondientes, lo que puede conducir a un sobre entrenamiento de los modelos si los datos son pocos. Para resolver este problema, Leung [58] propuso un enfoque de adaptación en que las distribuciones discretas de los modelos de locutores son una adaptación de las de modelos UBM. En su investigación Leung adoptó un enfoque similar al planteado por

Kirchhoff [55] para extraer los parámetros articulatorios correspondientes a "manner" y "place" de las articulaciones, reduciendo así la dimensión de entrada de los modelos de pronunciación, a sólo dos de las cinco propiedades de articulación sugeridas en su trabajo anterior [136]. Ellas fueron escogidas porque sus combinaciones son capaces de distinguir las consonantes y la mayoría de las vocales.

En este nuevo acercamiento las expresiones de entrenamiento de todos los locutores son utilizadas para formar un conjunto de modelos de fonemas para el CPM. El entrenamiento de los modelos se basó en etiquetas de fonemas convertidas a partir de transcripciones a nivel de palabras y de léxico, donde se aprobaron un total de 46 fonemas, incluyendo un silencio, un ruido de fondo, un ruido vocal y una risa. Para el entrenamiento de los modelos de fonemas y el reconocimiento se utilizaron vectores acústicos de 39 dimensiones, cada uno compuesto de 12 rasgos MFCC, la energía normalizada, y sus derivadas de primer y segundo orden. Cada uno de los 46 fonemas independientes del contexto, fue modelado por un HMM de tres estados de izquierda a derecha, con 16 GMM de covarianza diagonal por estado, utilizando HTK [137] para entrenar los HMM.

El software QuickNet [138] fue utilizado para entrenar los dos MLP, cada uno de los cuales se compone de 234 nodos de entrada (nueve tramas MFCC de dimensión 26: 12 MFCC, la energía y sus correspondientes coeficientes delta), 50 nodos ocultos y de 6 ó 10 nodos de salida en dependencia del parámetro articulatorio que se trate.

Los flujos de parámetros articulatorios alineados a las secuencias de fonemas de todos los locutores se utilizaron para formar un UBM representando las probabilidades de 60 combinaciones de clases "manner" y "place" condicionadas por 41 fonemas (excepto el silencio y el ruido) en el conjunto de fonemas establecido. La manera de obtener las alineaciones de los fonemas de las expresiones de entrenamiento fue coherente con las expresiones de verificación.

Este trabajo [58] abarcó dos enfoques para obtener un modelo de locutor basado en AF-CPM. En el primero, las probabilidades se calcularon sobre la base de los flujos de parámetros articulatorios y las secuencias de fonemas de cada locutor dado. En un segundo enfoque, las probabilidades de los locutores fueron adaptadas a partir del UBM. Estos resultados mostraron que existe información del locutor en los parámetros articulatorios y que la técnica propuesta AF-CPM es un método eficaz para modelar las variaciones de pronunciación de los hablantes. También se constató que el uso de flujos de parámetros articulatorios para CPM permite utilizar expresiones más cortas para el entrenamiento y la verificación.

7 Conclusiones

En esta área de investigación se han desarrollado diversos métodos de extracción de rasgos, cada uno orientado fundamentalmente a la búsqueda de un tipo de información específica. Una primera aproximación radica en la inclusión de la información acústica resultando los SDC los que mejores resultados han propiciado en este marco. Se quiere destacar que con este método se logra captar un poco de la dinámica natural del idioma por lo que se afirma que estos rasgos tienen un comportamiento seudo - prosódico.

Partiendo de los parámetros lingüísticos se encontró que los rasgos prosódicos han demostrado ser bastante discriminativos en términos de identificar idiomas. Son varios los métodos de extracción de rasgos que se usan para extraer información supra-segmental de la señal de voz. Los mejores resultados en este marco están dados por el método de extracción CWT [3]. La principal ventaja de este método radica en que tiene una muy buena resolución para las bajas frecuencias haciendo una separación entre estas y las altas frecuencias. Sin

embargo, queda la duda de si la especificidad con que trabaja este método afecta o no la capacidad discriminativa del conjunto de rasgos resultantes.

Por otra parte, resultó de gran interés el estudio de los parámetros articulatorios y las aplicaciones que se han hecho hasta el momento. Consideramos que la inclusión de este acercamiento en las tecnologías LID seria de gran beneficio, puesto que teóricamente las características articulatorias y su dinámica aportan un alto grado de información que podría resultar discriminativo para los idiomas. Los resultados obtenidos hasta el momento en esferas de reconocimiento multilingüe del habla, y reconocimiento del locutor nos alientan en este sentido.

En vista a lo antes expuesto se identificaron varios métodos de extracción de rasgos articulatorios, que podemos distribuir en 3 grandes grupos. El primer agrupamiento serían los métodos que se basan en la transcripciones acústicas - articulatorias, en este marco se encuentran algunos acercamientos que incluyen reconocimiento fonético como paso previo a la transcripción. El segundo grupo estaría dado por los métodos que implementan el mapeo o transformación articulatoria-acústica, ampliamente usados en aplicaciones de síntesis de voz. Y en el tercer grupo lo conformarían los métodos que intentan clasificar directamente a partir de las tramas del espectro las características articulatorias ("inverse-mapping").

Desde el punto de vista de la aplicación práctica consideramos que hasta el momento no se ha intentado su inclusión en un sistema automático de identificación de idiomas debido a que la obtención de los parámetros articulatorios puede volverse un tanto costosa en términos de disponibilidad de material de entrenamiento y del costo computacional.

La obtención de un método para la extracción de los parámetros articulatorios, incluyendo su dinámica, a partir del comportamiento acústico del habla, que permita discriminar entre los idiomas de una forma eficiente (lograr similar eficacia con menos datos y/o menos costo computacional) es un problema a resolver.

Referencias bibliográficas

[1] Y. K. Muthusamy*, et al.*, ""Perceptual benchmarks for automatic language identification" " in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'94)* Adelaide, Australia. , 2004, pp. pp 333-336. .

[2] E. Singer*, et al.*, ""Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition"," in *Proc. Eurospeech*, Geneva, Switzerland, ISCA, 1-4 September 2003, pp. pp. 1345-1348.

[3] A. L. Reyes, ""Un Método para la Identificación Automática del Lenguaje Hablado Basado en Características Suprasegmentales"," Tesis Doctoral, México, 2007.

[4] J. B. Bermúdez*, et al.*, *"Reconocimiento de voz y fonética acústica"*, 1ª, 1ª Reimpresión ed. Madrid, 2000.

[5] A. Quilis, *Tratado de fonología y fonética española.* Madrid, 2001.

[6] X. Huang*, et al.*, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.*: Pearson Education 2001.

[7] J. Benesty*, et al.*, *"Springer Handbook of Speech Processing"*. Berlin Heidelberg: Springer-Verlag, 2008.

[8] D. Cimarusti and R. B. Ives, "Development of an automatic identification system of spoken languages: Phase 1," in *ICASSP*, 1982, pp. pp. 1661-1663.

[9] P. A. Torres-Carrasquillo*, et al.*, " Language identification using Gaussian mixture model tokenization," in *ICASSP*, 2002, pp. pp. 757-760.

[10] M. A. Zissman, "Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models " in *ICASSP*, 1993, pp. pp. 399-402.

[11] K. K. Wong and M. Siu, " Automatic language identification using discrete hidden Markov model," presented at the Interspeech 2004.

[12] Y. K. Muthusamy, " *A Segmental Approach to Automatic Language Identification* ." Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, Beaverton 1993.

[13] W. M. Campbell*, et al.*, "Language recognition with support vector machines," in *Odyssey: The Speaker and Language Recognition Workshop* 2004, pp. pp. 41-44.

[14] T. J. Hazen, "Automatic Language Identification Using a Segment-Based Approach," Masters MIT, Cambridge 1993.

[15] T. J. Hazen and V. Zue, " Segment-based automatic language identification " *Acoust. Soc. Am.,* pp. pp 2323-2331, 1997.

[16] D. Crystal, *A Dictionary of Linguistics and Phonetics*, 2nd edn. ed. Oxford:: Basil Blackwell, 1985.

[17] P. Ladefoged, *Vowels and Consonants: An Introduction to the Sounds of Languages*. Oxford Blackwell, 2005.

[18] (2005, *Reproduction of the International Phonetic Alphabet, http://www.langsci.ucl.ac.uk/ipa/index.html*

[19] I. Maddieson, *Sound Patterns of Language.* Cambridge: Cambridge Univ. Press, 1984.

[20] A. Traill, *Phonetic and Phonological Studies of of !Xó˜o bushman.* Hamburg: *Quellen zur Khoisan- Forschung 1*, 1985.

[21] A. S. House and E. P. Neuberg, "Toward automatic identification of the languages of an utterance: preliminary methodological considerations.," *J. Acoust. Soc. Am,* pp. pp 708-713 1977.

[22] M. A. Zissman, " Comparison of four approaches to automatic language identification of telephone speech.," *IEEE Trans. Speech Audio Process.,* pp. pp 31-44, 1996.

[23] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Commun,* pp. pp (1-2), 31-51 2001.

[24] A. W. Black and T. Schultz, "Speaker clustering for multilingual synthesis," in *ISCA Tutorial and Research Workshop on Multilingual Speech and Language*, 2006.

[25] M. Adda-Decker*, et al.*, "Phonetic knowledge, phonotactics and perceptual validation for automatic language identification," presented at the International Congress of Phonetic Sciences, 2003.

[26] P. B. d. Mareüil*, et al.*, "Multi-lingual automatic phoneme clustering," presented at the Int. Congress Phonetic Sci, 1999.

[27] C. Corredor-Ardoy*, et al.*, " Language Identification with Languageindependent Acoustic Models.," in *Eurospeech*, 1997, pp. pp. 355-358.

[28] B. Ma and H. Li, "Spoken language identification using bag-of-sounds," in *International Conference on Chinese Computing*, 2005.

[29] K. Kirchhoff and S. Parandekar, "Multi-stream statistical language modeling with application to automatic language identification," in *7th European Conference on Speech Communication and Technology Proceedings of Eurospeech*, 2001, pp. pp. 803-806.

[30] J. Blevins, *The syllable in phonological theory. The Handbook of Phonological Theory* vol. 1. Oxford Blackwell Handbooks in Linguistics, 1995.

[31] M. Kenstowicz, *Phonology in Generative Grammar,* vol. 7. Oxford Blackwell Textbooks in Linguistics, 1994.

[32] D. Zhu*, et al.*, "Different size multilingual phone inventories and context-dependent acoustic models for language identification,," in *Interspeech* 2005, pp. pp. 2833-2836.

[33] K. Berkling*, et al.*, "Improving accent identification through knowledge of English syllable structure," in *5th International Conference on Spoken Language Processing*, 1998, pp. pp. 89-92.

[34] E. Grabe and E. L. Low, *Durational variability in speech and the rhythm class hypothesis*. Mouton de Gruyter, Berlin *Laboratory Phonology*, 2002.

[35] R. M. Dauer, *Stress-timing and syllable-timing reanalysed.*: J. Phonet, 1983.

[36] J. Rouas*, et al.*, " Modeling prosody for language identification on read and spontaneous speech," in *ICASSP*, 2003, pp. pp. 40-43,.

[37] J. Rouas, *et al.*, "Rhythmic unit extraction and modelling for automatic language identification.," *Speech Commun,* pp. pp 436-456, 2005.
[38] R. Tong, *et al.*, " Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *ICASSP* 2006, pp. pp. 205-208.
[39] T. Schultz, *et al.*, " Experiments with LVCSR based language identification," in *Speech Symposium SRS XV,* 1995.
[40] T. Schultz, *et al.*, " LVCSR-based language identification," in *ICASSP* 1996, pp. pp. 781- 784.
[41] D. Matrouf, *et al.*, "Language identification incorporating lexical information," in *5th International Conference on Spoken Language Processing*, 1998, pp. pp. 181-184.
[42] J. Hieronymus and S. Kadambe, "Robust spoken language identification using large vocabulary speech recognition," in *ICASSP* 1997, pp. pp. 779-782.
[43] C. Chelba, "*Exploiting Syntactic Structure for Natural Language Modeling* " Ph.D. thesis, Johns Hopkins University, Baltimore 2000.
[44] W. Wang and M. Harper, "The SuperARV language model: Investigating the effectiveness of tightly integrating multiple knowledge sources," in *Conference of Empirical Methods in Natural Language Processing*, 2002, pp. pp. 238-247.
[45] D. R. Gonzáles, "Rasgos dinámicos espectrales del habla y su relación con la prosodia en el reconocimiento de locutores," Telecomunicaciones y Electrónica, Istituto Superior Poliécnico Jose A. Echeverría (CUJAE), C. Habana, 2008.
[46] G. Miller and P. Nicely, "An analysis of some perceptual confusions among some English consonants," *JASA,* pp. pp 338-352, 1955.
[47] S. Singh, " Cross-language study of perceptual confusions of plosive phonemes in two conditions of distortion," *JASA,* pp. pp 635-656, 1966.
[48] M. D. Wang and R. C. Bilger., "Consonant confusions in noise: a study of perceptual features," *JASA,* pp. pp 1248-1266, 1973.
[49] R. W. Peters, "Dimension of perception for consonants," *JAcS* pp. pp 1985-1989, 1963.
[50] J. H. Greenberg and J. J. Jenkins, "Studies in the psychological correlates of the sound system of American English," *Word,* pp. pp 157-177, 1964.
[51] A. M. Liberman, *et al.*, "Perception of the speech code," *Psychological Review,* pp. pp 431-461, 1967.
[52] A. M. Liberman and I. G. Mattingly, " The Motor Theory of Speech Perception revised," *Cognition,* pp. pp 1-36, 1986.
[53] C. A. Fowler and M. R. Smith., "Speech perception as vector analysis. ," presented at the *Invariance and Variability in Speech Processes*, 1986.
[54] C. P. Browman and L. Goldstein., "Towards an articulatory phonology," *Phonology Yearbook 2,* pp. pp 219-252, 1986.
[55] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information," Doktor–Ingenieur, Technischen Fakult, Universit¨at Bielefeld, 1999.
[56] K. Erler and L. Deng, "Hidden Markov model representation of quantized articulatory features for speech recognition," *Computer speech & language ISSN 0885-2308* vol. vol. 7, no3,, pp. pp. 265-282 1993.
[57] K. Y. Leung, *et al.*, " Articulatory feature-based conditional pronunciation modeling for speaker verification," in *ICSLP*, 2004, pp. pp. 516-519.
[58] K. Y. Leung, *et al.*, "Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification," *Science Direct Speech Communication,* vol. 48, pp. pp 71-84, 2006.
[59] *S. Parandekar* and *K. Kirchhoff*, "Multi-stream language identification using data-driven dependency selection," presented at the Acoustics, Speech, and Signal Processing, 2003., 2003.
[60] S. Kanokphara and J. Carson-Berndsen, "Articulatory-Acoustic-Feature-based Automatic Language Identification," in *ISCA Workshop on Multilingual Speech and Language Processing (MULTILING 2006),* Stellenbosch, South Africa, 2006.
[61] C. P. Browman and L. Goldstein., "Articulatory phonology: an overview," *Phonetica,* pp. pp 155-180, 1992.

[62] R. P. Cohn., "Robust voiced/unvoiced speech classification using a neural net," presented at the *Proceedings of the Acoustics, Speech, and Signal Processing -91*, Washington, DC, USA 1991.

[63] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals," presented at the 6th International Congress on Acoustics Tokyo, Japan, 1968.

[64] J. D. Markely and A. H. Gray, *Linear Prediction of Speech*. Berlín: Springer Verlag, 1976.

[65] B. S. Atal, "*Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.*," *Journal of the Acoustical Society of America,* p. pág. 55, 1974.

[66] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, Englewood Cliffs, NJ: Prentice-Hall Signal Proc Series, 1978.

[67] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. New York: Marcel Dekker, Inc., 2001.

[68] G. H. Sierra, "*Extracción de rasgos acústicos.*," Departamento de Reconocimiento de patrones, Cenatav., La Habana 2006.

[69] R. Cole*, et al.*, "Survey of the State of the Art in Human Language Technology.," Cambridge University1997.

[70] B. P. Bogert*, et al.*, " *The quefrency analysis of time series for echoes:Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking.*," in *Symposium on Time Series Analysis*, New York: Wiley, 1963, pp. Chapter 15, 209-243.

[71] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions,* vol. 28, pp. pp 357 - 366, 1980.

[72] R. J. Mammone*, et al.*, ""Robust speaker recognition: a feature-based approach,"" *Signal Processing Magazine, IEEE,* vol. 13, pp. pp. 71 - 129, 1996.

[73] J. Pelecanos*, et al.*, " Enhancing automatic speaker identification using phoneme clustering and frame based parameter and frame size selection," presented at the *Fifth International Symposium on Signal Processing and Its Applications*, 1999.

[74] D. A. Reynolds, ""Experimental evaluation of features for robust speaker identification," " *IEEE Transactions on Speech and Audio Processing,* vol. 2, pp. 639{643,, 1994.

[75] R. Lawrence and J. B.-. Hwang, "*Fundamentals of Speech Recognition*". New jersey: Prentice Hall, 1993.

[76] P. A. Torres-Carrasquillo*, et al.*, ""Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features"," in *International Conference on Spoken Language Processing*, Denver, 2002, pp. pp. 33-36, 82-92.

[77] B. Bielefeld, ""*Language identification using shifted delta cepstrum"* " presented at the Fourteenth Annual Speech Research Symposium, 1994.

[78] F. Allen, "*Automatic Language Identification*," PhD Thesis., University of New South Wales, Sydney, 2005.

[79] A. Cutler*, et al.*, "Prosody in the comprehension of spoken language: A literature review," *Language and Speech,* vol. 40, pp. pp 141-202

[80] *L. Mary* and *B. Yegnanarayana*, "*Prosodic features for speaker verification*," presented at the *INTERSPEECH 2006 - ICSLP*, 2006.

[81] C.-Y. Lin and H.-C. Wang, "Language Identification Using Pitch Contour Information,," in *Acoustics, Speech, and Signal Processing, ICASSP* 2005, pp. pp 601 - 604.

[82] Y. Obuchi and N. Sato, "Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization," in *Acoustics, Speech, and Signal Processing, ICASSP* 2005, pp. pp 569 - 572.

[83] F. Cummins*, et al.*, "Language Identification from Prosody without explicit Features," in *EUROSPEECH'99*, Budapest, Hungary,, 1999, pp. pp. 371-374.

[84] A. Samouelian, ""Automatic Language Identification using Inductive Inference"," presented at the 4th International Conference on Spoken Language Processing ICSLP Philadelphia, USA., 1996.

[85] R. Modic, *et al.*, "Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English speechDat2"," presented at the NOLISP, Le Croisic, France, 2003.
[86] M. Gupta and A. Gilbert, ""Robust speech recognition using wavelet coefficient features"," presented at the IEEE Automatic Speech Recognition and Understanding Workshop, USA, 2001.
[87] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," presented at the Speech Communication, Amsterdam, The Netherlands, 1991
[88] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Acoustics, Speech, and Signal Processing ICASSP* 2003, pp. pp 68-71.
[89] R. M. Hegde, *et al.*, " Application of the Modified Group Delay Function to Speaker Identification and Discrimination," in *ICASSP*, 2004, pp. pp. 517-520.
[90] R. M. Hegde, *et al.*, "Continuous Speech Recognition using Joint Features derived from The Modified Group Delay Function and MFCC," presented at the INTERSPEECH-ICSLP, 2004.
[91] "Database for Indian languages," ed. Chennai, India: Speech and Vision Lab, IIT Madras, 2001.
[92] Y. K. Muthusamy, *et al.*, "The OGI multilanguage telephone speech corpus," in *ICSLP*, 1992, pp. pp. 895-898.
[93] R. M. Hedge and H. A. Murthy, "Automatic Language Identification and discrimination using the modified group delay feature," in *Conf. on Intelligent Sensing and Information Processing*, Chennai, 2005, pp. pp. 395-399.
[94] S. Greenberg and T. Arai, "What are the Essential Cues for Understanding Spoken Languages?," *IEICE Transaction on Information & System,* vol. E87-D, p. pp. 1059, 2004.
[95] V. Dimitriadis, *et al.*, "Robust AM-FM Features for Speech Recognition," *IEEE Signal Processing Letters,* vol. 12, pp. pp. 621-624, 2005.
[96] D. Dimitriadis and P. Maragos, "Robust Energy Demodulation Based on Continuous Models with Application to Speech Recognition.," in *Eurospeech-03*,, Geneva, 2003.
[97] P. Maragos, *et al.*, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing,* vol. 41, pp. pp. 3024-3051, 1993.
[98] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing Magazine, IEEE,* vol. 37, pp. pp. 95-120, 1994.
[99] T. Thiruvaran, *et al.*, "Extraction of FM components from speech signals using all-pole model," *Electronics Letters,* vol. 44, pp. pp. 449-450.
[100] T. Yin, *et al.*, "Introducing a FM based Feature to Hierarchical Language Identification," presented at the InterSpeech 08, 2008.
[101] F. Metze and A. Waibe, "A flexible stream architecture for ASR using articulatory features," in *International Conference on Spoken Language Processing*, 2002, pp. pp. 2133-2136.
[102] M. Tang, *et al.*, " Modeling linguistic features in speech recognition," in *European Conference on Speech Communication and Technology*, 2003, pp. pp. 2585-2588.
[103] T. A. Stephenson, *et al.*, "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables," in *International Conference on Spoken Language Processing*, 2000, pp. pp. 951-954.
[104] M. Richardson, *et al.*, "Hidden-articulator markov models: Performance improvements and robustness to noise," *International Conference on Spoken Language Processing,* vol. III, pp. pp. 131-134, 2000.
[105] J. Frankel and S. King, "ASR - articulatory speech recognition," in *European Conference on Speech Communication and Technology*, 2001, pp. pp. 599-602.
[106] K. Y. Leung and M. Siu, "Speech recognition using combined acoustic and articulatory information with retraining of acoustic model parameters," in *International Conference on Spoken Language Processing*, 2002, pp. pp. 2117-2120.
[107] L. Zhang and W. Edmondson, "Speech recognition based on syllable recovery," in *European Conference on Speech Communication and Technology*, 2003, pp. pp. 2537-2540.
[108] K. Kirchhoff, *et al.*, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication,* vol. 37, pp. pp 303-319, 2002.

[109] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, ed New York: Marcel Dekker, 1992, pp. pp. 231-267.

[110] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.* , vol. 2, pp. pp 133-150, 1994.

[111] O. Schmidbauer, "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations," in *ICASSP-89*, 1989, pp. pp 616-619.

[112] P. Steingrimsson*, et al.*, "From acoustic signal to phonetic features: a dynamically constrained self-organising neural network," in *International Congress of Phonetic Sciences*, 1995.

[113] P. Dalsgaard, "Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions," *Computer speech & language,* vol. 6, pp. pp 303-329, 1992.

[114] L. Deng and K. Erler, "Microstructural speech units and their HMM representations for discrete utterance speech recognition," in *ICASSP-91*, 1991, pp. pp 193-196.

[115] L. Deng and D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulator features for all classes of English sounds," in *ICASSP-94*, 1994, pp. pp 45-47.

[116] K.-Y. Leung and M. Siu, "Articulatory-feature-based confidence measures," *Computer Speech and Language,* 2005.

[117] *K. Kirchhoff* and *S. Parandekar*, "Multi-Stream Statistical N-Gram Modeling With Application To Automatic Language Identification," in *Eurospeech*, Scandinavia, 2001.

[118] S. Kanokphara and J. Carson-Berndsen, "Better HMMBased Articulatory Feature Extraction with Context- Dependent Model," in *18th International Florida Artificial Intelligence Research Society Conference*, 2004.

[119] R. Jakobson*, et al.*, "Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates. ," *MIT Press,* 1969.

[120] S. Kanokphara*, et al.*, "Comparative Study: HMM&SVM for Automatic Articulatory Feature Extraction," in *9th Int'l. Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems*, 2006.

[121] Y. Shiga and S. King, "Accurate spectral envelope estimation for articulation-to-speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. pp. 19-24.

[122] T. Kaburagi and M. Honda, " Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory- acoustic database.," in *ICSLP-94*, Sydney, Australia, 1998, pp. pp. 433- 436.

[123] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.,* pp. pp 175-185, 2004.

[124] K. Nakamura*, et al.*, "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum," in *ICASSP.* , 2006, pp. pp. 93-96.

[125] C. T. Kello and D. C. Plaut, " A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters.," *J. Acoust. Soc. Amer,* pp. pp 2354-2364., 2004.

[126] J. Hogden*, et al.*, "Accurate recovery of articulator positions from acoustics: new conclusions based on human data," *J. Acoust. Soc. Amer,* pp. pp 1819-1834., 1996.

[127] S. Suzuki*, et al.*, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," in *ICSLP*, Sydney, Australia, 1998, pp. pp. 2251- 2254.

[128] K. Richmond*, et al.*, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech Language 17 (2),* pp. pp 153-172, 2003.

[129] T. Toda*, et al.*, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Science Direct Speech Communication* vol. 50, pp. pp 215-227, 2008.

[130] Y. Stylianou*, et al.*, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.6 (2),* pp. pp 131-142, 1998.

[131] A. Wrench. (1999, The MOCHA-TIMIT articulatory database http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.
[132] M. Adda-Decker and L. Lamel, "Pronunciation variants across systems, languages, and speaking style," in *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition* Kerkrade, Netherlands, 1998, pp. pp 1-6.
[133] E. Fosler-Lussier, "A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition," in *Text- and Speech-Triggered Information Access*. vol. 2705, ed Columbia University, New York: Springer Berlin / Heidelberg, 2003, pp. pp 38-77.
[134] D. Reynolds*, et al.*, "The superSID project: exploiting high-level information for high-accuracy speaker recognition.," in *ICASSP*, Hong Kong, 2003, pp. pp. 784-787.
[135] D. Klusác˘ek*, et al.*, "Conditional pronunciation modeling in speaker detection," in *ICASSP* 2003, pp. pp. 804-807.
[136] K. Y. Leung*, et al.*, " Applying articulatory features to telephone-based speaker verification.," in *ICASSP*, Montreal, 2004, pp. pp. 85–88.
[137] S. Young*, et al.*, *The HTK book for HTK 3.0. Tech*: Rep., Microsoft Corporation, 2000.
[138] P. Farber, "Quicknet on multispert: fast parallel neural network training.," ICSI1997.

RT_036, octubre 2010
Aprobado por el Consejo Científico CENATAV
Derechos Reservados © CENATAV 2010
Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
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