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

Reconocimiento del idioma hablado: tendencias actuales

Ana Montalvo Bereau y
José R. Calvo de Lara

RT_057 **octubre 2013**

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

Reconocimiento del idioma hablado: tendencias actuales

Ana Montalvo Bereau y
José R. Calvo de Lara

RT_057 **octubre 2013**

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

Tabla de contenido

Reconocimiento del idioma hablado: tendencias actuales

Ana Montalvo Bereau y José R. Calvo de Lara

Dpto. Reconocimiento de Patrones, Centro de Aplicaciones de Tecnologías de Avanzada(CENATAV),
La Habana, Cuba
{amontalvo, jcalvo}@cenatav.co.cu

RT_057, Serie Azul, CENATAV
Aceptado: 22 de octubre de 2013

Resumen. El presente reporte constituye una actualización de los Reportes Ténicos 015 de Noviembre del 2009 y 036 de Octubre de 2010. Los mismos versan sobre la metodología para el reconocimiento automático del idioma hablado, por lo que el presente no se centrará en la descripción profunda de las técnicas establecidas hasta entonces, sino que dará seguimiento a la evolución de los temas investigativos más fuertes y promisorios, así cómo explorará nuevas ramas de interés.

Palabras clave: reconocimiento del idioma hablado, rasgos acústicos, fonotácticos, *tokenización*.

Abstract. This report is an update of the techs Reports 015 and 036 of November 2009 and October 2010. They relate to the methodology for the automatic spoken language recognition, so that this report will not focus in-depth description of the previously established techniques, but it will follow the evolution of more strong and promising's research topics , and also explore new areas of interest.

Keywords: spoken language recognition, acoustic features, tokenization, phonotactic features.

1 Introducción

El habla es la manifestación acústica del lenguaje, y es probablemente la principal forma de comunicación entre humanos. El desarrollo de las telecomunicaciones y del procesamiento digital de la información ha demandado esfuerzos por comprender los mecanismos de comunicación mediante habla.

Entre las numerosas aplicaciones que abarca el campo de análisis de la señal de habla están:

- codificación de la señal: compresión de la información contenida en la señal para almacenar o mejorar la velocidad de transmisión.
- síntesis: generación automática de habla a partir de un texto arbitrario.
- procesamiento de la señal: inteligibilidad, descontaminación, encriptado, cancelación de eco.
- reconocimiento y comprensión del habla, reconocimiento del locutor y del idioma.

Es a esta última tarea de reconocimiento que está dedicada esta investigación. El reconocimiento del idioma hablado consiste en determinar el idioma en que se habla basándose sólo en una

muestra de voz, sin considerar al hablante o la semántica de lo que está diciendo. El reconocimiento automático del idioma hablado (SLR[1]), de manera general, es el proceso por el cual el idioma de una muestra de señal de voz digitalizada es reconocida por una computadora.

En la actualidad existe un conjunto grande de sistemas cuyos resultados finales dependen en gran medida de la tecnología SLR, por lo que es un módulo importante dentro de varias aplicaciones como:

- sistemas de conversación multilingüe,
- traducción de idioma hablado,
- reconocimiento del habla multilingüe,
- recuperación de documentos hablados.

También es un asunto de gran importancia en las áreas de inteligencia y seguridad donde se necesita conocer el idioma de mensajes grabados y documentos archivados antes de poder extraer cualquier información de ellos. Otra aplicación en esta rama sería la determinación del origen del hablante a partir del reconocimiento de su lenguaje autóctono.

Adaptar los sistemas de procesamiento del habla a un nuevo idioma es actualmente uno de los retos del desarrollo de las tecnologías del habla multilenguaje, en particular es un desafío al que se enfrenta también nuestro grupo de trabajo con varios proyectos, los cuales hasta el momento no son capaces de dar una respuesta en correspondencia con el idioma de las señales que se procesan.

Por tanto, se requieren nuevos modelos y algoritmos que permitan el reconocimiento del idioma, para que los sistemas sean capaces de generar una salida consecuente con el idioma de la entrada.

2 Principios del reconocimiento del idioma hablado

Los seres humanos nacen con la habilidad de discriminar idiomas hablados como parte de su inteligencia [1]. Al igual que cualquier otra tarea de inteligencia artificial, el SLR busca replicar dicha habilidad humana por medios computacionales. El asunto clave es cómo medir científicamente la individualidad de cada idioma existente en el mundo.

Hace más de una década que el SLR dejó de ser ciencia ficción, siendo llevado a la práctica en numerosas aplicaciones [2] [3]. Se estima que existen varios miles de idiomas en el mundo [4,5]. La reciente edición de "Ethnologue", base de datos que describe todos los idiomas [6], ha documentado 6909 idiomas hablados, latentes.

Por otra parte, el reconocimiento de idioma basado en texto es formulado en la actualidad como un problema de categorización de texto [7], y para los idiomas que emplean el alfabeto latino, se han alcanzado rendimientos tan buenos que es considerado un problema resuelto [8]. En la práctica, el SLR constituye un reto mucho mayor que su homólogo basado en texto, ya que no existe garantía de que una máquina sea capaz de transcribir habla sin errores.

Es sabido que los humanos reconocen los idiomas a través de un proceso psicoacústico perceptual, inherente al sistema auditivo, y es esta la fuente de inspiración para el SLR [1].

[1] *Spoken Language Recognition*

2.1 Caracterización de idiomas

Para identificar un idioma, así como para resolver cualquier problema de reconocimiento de patrones, es necesario escoger una representación adecuada de las características discriminativas que nos interesan explotar. ¿Donde la señal de voz porta la información de interés para reconocer el idioma?¿Qué nivel de información nos resulta más útil para esta tarea particular?

El primer experimento perceptual para medir cuán bien los oyentes podían reconocer un idioma hablado, fue reportado en [9]. El mismo concluía que las personas, con adecuado entrenamiento, eran el reconocedor más eficaz, observación que se mantiene luego de 15 años [10], dado que las personas evaluadas tenían dominio léxico y semántico de los idiomas. Por otra parte para los idiomas con los cuales no tenían familiaridad, las personas eran capaces de hacer juicios subjetivos con referencias a idiomas que dominaban, por ejemplo: "...suena como árabe", "...es tonal como el mandarín o el vietnamita", o "... tiene patrones de acentuación como el inglés o el alemán".

Aunque estos juicios eran menos precisos, mostraban cómo los oyentes aplican su conocimiento lingüistico a diferentes niveles para distinguir determinado grupo de idiomas.

Otros experimentos también han mostrado que el ser humano, luego de determinado tiempo expuesto a un idioma del cual no tiene ningún conocimiento lingüístico, es capaz de reconocerlo de otros igualmente desconocidos para él. O sea, se percata de cierta información contenida en el habla, y no precisamente en las palabras, que le permite identificar el idioma escuchado.

El conjunto de pistas o claves del idioma contenido en el habla, es ilustrado en Fig.1 de acuerdo a su nivel de abstracción del conocimiento.

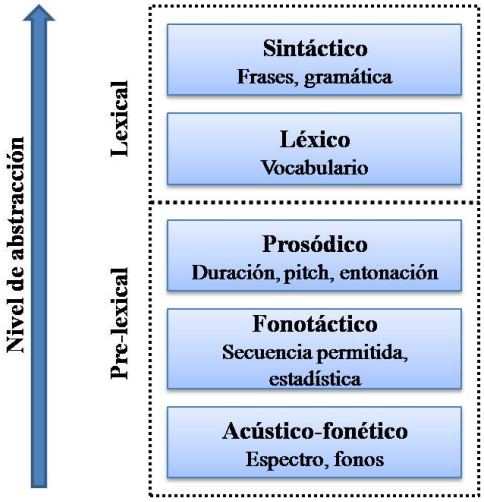

Fig. 1. Niveles de información contenida en el habla para el *SLR*.

Acústico-fonético El aparato fonador humano es capaz de producir un amplio rango de sonidos. Los sonidos del habla, como eventos acústicos concretos, son llamados fonos; mientras que vistos como entidades de un sistema lingüístico, son referidos como fonemas [11]. El número de fonemas empleados en un idioma varía entre 15 y 50, teniendo la mayoría alrededor de 30 fonemas cada uno. Existen diferencias entre los repertorios fonéticos de los diversos idiomas, de hecho cada idioma tiene su conjunto propio de fonemas, por tanto su propia distribución de rasgos acústicos-fonéticos.

Fonotáctico Existe un conjunto de reglas, denominadas fonotácticas, que dictan las secuencias permitidas de fonemas en un idioma. Los fonemas pueden ser compartidos por varios idiomas, pero la estadística de sus patrones secuenciales es exclusiva de cada uno.

Prosodia Esta engloba a un grupo de carácterísticas suprasegmentales tales como el estrés, la duración, el ritmo y la entonación. La prosodia ha probado ser muy útil para distinguir entre grupos de idiomas (tonal, no tonal), sin embargo los experimentos auditivos en humanos muestran que los rasgos porsódicos aportan mucho menos información discriminativa para los idiomas que los fonotácticos. No obstante sigue siendo un reto extraer información prosódica confiable e incorporarla con éxito a la tarea SLR.

Léxica y sintáctica La información léxica está relacionada con el vocabulario de cada idioma, y aunque es algo bastante propio de cada idioma y que permite a los sistemas discriminar entre los idiomas candidatos, para el idioma hablado es una información generalmente costosa de obtener. La sintaxis por su parte, es la forma en que se organizan la palabras para formar oraciones o frases.

El uso de la informacón fonética y fonotáctica toma sentido al asumir que los idiomas poseen conjuntos parcialmente solapados de fonemas. Dicha hipótesis es la que sirvió de base para construir el alfabeto fonético internacional (IPA[2]) que muestra que aunque hayan 6909 idiomas en el mundo, el número total de fonos que se requieren para representar todos los sonidos está alrededor de los 300 [12].

2.2 Formulación del problema de reconocimiento de idioma hablado

Sea X una representación de la señal de voz, el problema de identificar el idioma pudiera formularse:

$$L^* = argmax_{L \in \mathcal{L}} P(L|X), \tag{1}$$

donde L^* es el idioma identificado del conjunto $\mathcal{L}$ de potenciales idiomas.

Aplicando Bayes tenemos que:

$$L^* = argmax_L P(X|L) * P(L), \tag{2}$$

donde $P(X|L)$ es la probabilidad de que X sea cierta dado el idioma L y $P(L)$ es la probabilidad a priori de L.

Asumiendo, sin pérdida de generalidad, que todos los idiomas son equiprobables, 2 pudiera simplificarse a:

$$L^* = argmax_L P(X|L). \tag{3}$$

De acuerdo a Fig.2, durante el procesamiento acústico se parametriza la señal de entrada, obteniéndose una cadena de rasgos.

Estos son procesados por el decodificador de habla, el cual, en correspondencia con la aproximación a emplear produce una secuencia de *tokens* (fonemas, sílabas, palabras, índices) o una representación acústica más compleja como super vectores (SVs).

La salida del decodificador de habla se enfrenta a los modelos del lenguaje (LMs[3]) previamente creados y se toma una decisión. El idioma asociado al modelo más probable es seleccionado como el idioma de la muestra de prueba en concordancia con 1.

[2] *International Phonetic Alphabet*
[3] *Language Models*

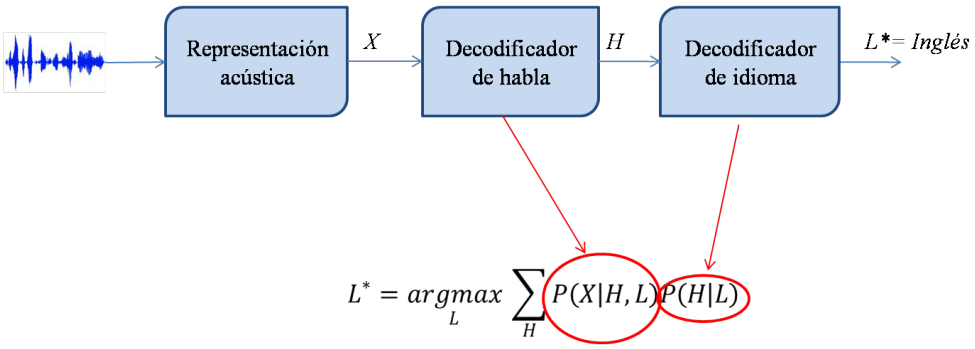

Fig. 2. Modelo del problema SLR.

De forma más general puede observarse en la Fig.2 que a la probabilidad del lenguaje L^* contribuyen: la probabilidad de la secuencia del decodificador de habla dada la observación y la probabilidad del idioma dada la secuencia H.

2.3 Tendencias actuales

Un amplio espectro de enfoques al problema SLR ha sido propuesto. Sin embargo siguen dominando el área los dos paradigmas tradicionales, que pudieran definirse por los rasgos que empleen para modelar el idioma:

- los que emplean rasgos acústicos de bajo nivel (**Enfoque acústico-fonético** basado en las características espectrales *short-time* de la señal de audio).
- aquellos que usan rasgos fonotácticos de alto nivel (**Enfoque fonotáctico** que usa la estadística presente en la secuencia de *tokens*).

En cuanto a los recursos que demanda cada aproximación, la acústica solo requiere de las grabaciones de audio para entrenar al sistema, mientras que la aproximación fonotáctica requiere audio etiquetado fonéticamente para la creación de los modelos acústicos de los *tokens* y grandes corpus de texto para la creación de los modelos del idioma. Por otra parte la fonotáctica es computacionalmente mucho más costosa, pudiendo tardar decenas de veces más que los algoritmos de reconocimiento acústicos [13]. Todas las aplicaciones en la actualidad fusionan amabas aproximaciones, ya que brindan información complementaria, no obstante la fonotáctica es la más eficaz como método *stand-alone*, sin desestimar que para señales cortas la acústica tiene un mejor desempeño [14].

En los trabajos enviados al NIST-LRE[4] 2011, llama la atención como en todos los sistemas subscritos se fusionan ambos paradigmas y en muchos casos los subsistemas que los forman son desarrollados por distintos laboratorios. Es de remarcar también, los grandes volúmenes de datos empleados para el entrenamiento (centenares de horas) y como la *tokenización* es llevada a cabo en casi la totalidad de los casos por reconocedores fonéticos (ver Tabla 1).

En el presente reporte abordaremos los avances que tuvieron lugar dentro de estas dos líneas investigativas, partiendo de una breve introducción que asume un conocimiento previo de las bases teóricas de ambas [15,16].

[4] *National Istitute of Standarization Technologies - Language Recognition Evaluation*

Tabla 1. Algunos sistemas presentados en NIST-LRE 2011.

Título del trabajo	Subsistemas (**Acústicos** - Fonotácticos)
UTD-CRSS SYSTEMS FOR NIST LANGUAGE RECOGNITION EVALUATION 2011. University of Texas at Dallas Center for Robust Speech Systems	- **i-vector System** - **SVM-GSV System** - PPRLM System - Combined Articulatory and Prosody System
Description and analysis of the Brno276 system for LRE2011 Brno University of Technology (BUT), Politecnico di Torino (PoliTo) and AGNITIO. (2000 horas para los subsistemas fonotácticos)	- **i-vector-2048FG (acoustic i-vector extractor)** - PHN-HU-i-vector (phonotactic i-vector extractor) - PHN-RU-PCA (PCA) - PHN-ENG-BT (binary decision tree)
University of the Basque Country (EHU) Systems for the 2011 NIST LRE. (437 horas de entrenamiento)	- **Dot-Scoring** - **i-vector** - Phone-SVM-CZ - Phone-SVM-HU - Phone-SVM-RU
The L2F Language Recognition System for NIST LRE 2011 Spoken Language Systems Lab, INESC-ID Lisboa, Portugal. 60 horas (en promedio, menos de 2.5 horas por idioma)	- PRSVM-pt - PRSVM-br - PRSVM-es - PRSVM-en GSV - **i-Vector**
The MITLL NIST LRE 2011 Language Recognition System Massachusetts Institute of Technology Lincoln Laboratory (867 horas para el entrenamiento)	- **GMM-MMI** - **SVM-GSV** - **i-vector systems** - PRSVM token system
IIR System Description for the 2011 NIST LRE Institute for Infocomm Research, Singapore (52 horas para el entrenamiento)	- **KL-SVM** - **BHATT-SVM** - **IV-400** - TALR

3 Enfoque fonotáctico

Este enfoque está motivado por la hipótesis de que los idiomas hablados pueden modelarse teniendo en cuenta las ligaduras léxicas-fonológicas (2.1). Los sistemas basados en este enfoque son llamados también de alto nivel, pues explotan tres niveles de información: la acústica (parte de rasgos acústicos), la fonética (recordar que los fonemas son abstracciones formales de los sonidos del habla, no tienen identidad física y dependen del lenguaje) y la fonotáctica (reglas que rigen la distribución de los sonidos fonéticos dentro de un lenguaje).

El primer intento de utilizar los patrones fonotácticos para reconocer idiomas, fue comparando la frecuencia de ocurrencia de determinados sonidos de referencia en la muestra de prueba, con frecuencia de los mismos sonidos en las muestras de idioma conocido. Para esto se necesita primeramente *tokenizar* la señal en las unidades sonoras especificadas, donde llamaremos *tokenizar* al proceso de obtención de los *tokens*, que no son más que unidades estructurales (fonemas, eventos acústicos, mezclas gaussianas) empleadas para representar el discurso hablado.

El enfoque fonotáctico consta de dos etapas principales: la *tokenización* y el modelado del lenguaje.

Esta aproximación asume al habla como una fuente de *tokens* que representan el discurso, por lo que el primer paso luego de la parametrización acústica es representar la señal con dichas unidades. Este flujo de *tokens* es usado para extraer características y entrenar clasificadores (modelos del idioma), para durante la etapa de prueba llevar a cabo el proceso de clasificación.

Se verá más adelante como a diferencia del enfoque acústico, los sistemas basados en la aproximación fonotáctica sí requieren una importante cantidad de información específica del idioma.

3.1 Tokenización

Un tokenizador de habla convierte la señal en una secuencia de símbolos. Estos *tokens* describen un determinado atributo acústico y pueden ser de distinto tamaño, yendo desde una trama (10ms), un fonema, una sílaba, a incluso una palabra.

La técnica de *tokenización* empleando modelos de mezclas gaussianas (GMMs[5]) [17] opera sobre los rasgos, al nivel de las tramas. Convierte una secuencia de tramas de habla en una secuencia de etiquetas de Gaussianas, cada una de las cuáles maximiza la probabilidad de la correspondiente trama.

Como este proceso no asocia el evento acústico con un fonema, no requiere de etiquetado alguno para entrenar el modelo, lo cual constituye un aspecto de peso que sigue atrayendo a muchos invesdtigadores [18]. Desafortunadamente, este análisis solo captura la dinámica a un alcance de decenas de milisegundos, lo cual es un intervalo muy corto para capturar la información fonológica, no obstante es un reto lograr extraer la información, sin dudas presente a ese nivel, con un menor costo (como lo constituyen los etiquetados).

De manera general, la *tokenización* empleando GMMs tiene menor eficacia que los métodos para el reconocimiento fonético basados en cadenas ocultas de Markov (HMM[6]) o redes neuronales (ANN[7]) en el reconocimiento de idioma hablado.

Otro método de *tokenización* es el que usa atributos articulatorios [19,20,18]. En este caso los *tokens* pertenecen a un conjunto común para todos los idiomas, lo cuál es una importante

[5] *Gassian Mixture Models*
[6] *Hidden Markov Models*
[7] *Artificial Neural Networks*

ventaja, ya que los datos disponibles para distintos idiomas pueden ser compartidos y empleados para construir un reconocedor universal de atributos (UAR[8]).

Un objetivo similar se persigue en [21], con la importante diferencia de que UAR usa un conjunto mucho menor de unidades (15 atributos articulatorios) y de que necesariamente parte de un etiquetado articulatorio de los fonos para entrenar el sistema, cosa que no se requiere para los ASM[9] de [21].

Como un compromiso entre costo de desarrollo y efectividad, los fonemas son los *tokens* más empleados en los sistemas del estado del arte. Los reconocedores fonéticos (PR[10]) son típicamente modelados con HMMs.

Se ha observado en la práctica que un PR entrenado para un idioma, es capaz de *tokenizar* cualquier otro idioma.

Otro ejemplo que ha sido de gran impacto en la comunidad científica es el PR desarrollado por la universidad checa de Brno [22]. Dicho sistema se basa en la estructura TRAP[11] propuesta en [23] cuya mayor ventaja es la posibilidad de extraer información contextual de un intervalo de tiempo mayor (300-400ms). A su vez dedica gran parte de la investigación a la obtención de un híbrido ANN-HMM.

A la hora de identificar el idioma de una muestra desconocida, se parte de la secuencia de *tokens* que fue obtenida por el *tokenizador* y se compara con los modelos de los lenguajes previamente creados durante el entrenamiento del sistema.

3.2 Modelado del idioma basado en n-gramas de fonemas

La utilización de enfoques probabilísticos para el modelado del idioma ha tomado un nuevo impulso en los últimos años, sin duda provocado por la disponibilidad de grandes cantidades de información en formato digital.

Los modelos del idioma tienen por objetivo calcular, dada una secuencia de *tokens*, su probabilidad de aparición en un idioma. Para resolver este problema, los enfoques probabilistas no utilizan conocimiento lingüístico profundo, sino que parten de la información obtenida a partir de corpus de gran tamaño que suponen representativos del idioma y, utilizando el principio de máxima verosimilitud, asignan las probabilidades correspondientes según las frecuencias de aparición en el corpus. Por supuesto, sobre esta base existen diversas elaboraciones que pueden incorporar en mayor o menor medida información lingüística o de otro tipo para afinar los resultados.

¿Qué son los n-gramas?

Los LM estiman la probabilidad de una secuencia de *tokens*, en nuestro caso secuencia de fonemas $\hat{P}(p_1, p_2, ..., p_m)$, lo que significa que evalúan $P(p_i)$ de acuerdo a Bayes:

$$P(p_i|\mathcal{O}) = \frac{P(\mathcal{O}|p_i) * P(p_i)}{P(\mathcal{O})}, \tag{4}$$

donde p_i representa el i-ésimo fonema y $\mathcal{O} = o_1, o_2, ..., o_T$ la secuencia de vectores observados. La probabilidad $\hat{P}(p_1, p_2, ..., p_m)$ puede ser escrita como un producto de probabilidades condicionales:

$$\hat{P}(p_1, p_2, ..., p_m) = \prod_{i=1}^{m} \hat{P}(p_i|p_1, ..., p_{i-1}). \tag{5}$$

[8] *Universal Attribute Recognition*
[9] *Acoustic Segment Model*
[10] *Phone Recognizers*
[11] *Temporal Patterns*

La ecuación 5 brinda la posibilidad de aproximar $\hat{P}(p_1, p_2, ..., p_m)$ limitando el contexto:

$$\hat{P}(p_1, p_2, ..., p_m) \simeq \prod_{i=1}^{m} \hat{P}(p_i | p_{i-n+1}, ..., p_{i-1}), \tag{6}$$

para $n \geq 1$. Si partimos del presupuesto de que el idioma es ergódico [24], o sea que tiene la propiedad de que la probabilidad de cualquier estado puede ser estimado de un volumen suficientemente grande de historia o información precedente, entonces la ec.6 es exacta. Lo anterior significa que es posible determinar la probabilidad de ocurrencia de un fonema dado un número $n-1$ de fonemas precedentes. Los modelos que emplean contexto limitado de esta forma, son conocidos como LM de n-gramas.

El valor de n está típicamente entre 1 y 4, y la componente de contexto condicional de la probabilidad ("$p_{i-n+1}, ..., p_{i-1}$" en ec.6) es llamada "historia".

La elección del valor de n tiene un peso importante en el número de parámetros potenciales que el modelo pudiera tener, el cuál es acotado superiormente por $|\mathcal{P}^n|$ donde $\mathcal{P}$ es el conjunto de fonemas del idioma (o repertorio fonético). Sin embargo no es solo el aspecto del espacio de almacenamiento el que tiene que considerarse debido a esta relación potencial. Se requieren grandes cantidades de texto de entrenamiento para asegurar significancia estadística de cada uno de los n-gramas y conferirle confiabilidad al modelo.

Resulta entonces que los n-gramas son secuencias de símbolos sobre las que se apoya el LM para predecir un símbolo dados sus $n-1$ predecesores.

Volvamos entonces al LM basado en n-gramas de fonemas. Este método usualmente utiliza la salida de uno o varios PRs como corpus de entrenamiento. Los sistemas con esta configuración son denotados PRLM o PPRLM[12]. En dichos sistemas, el (los) PR (s) *tokeniza* (n) la señal basado (s) en un inventario fonético común, mientras que los LMs describen la estructura de cada idioma en términos de estadística de los fonemas, de acuerdo al número escogido de gramas.

Para ilustrar cómo funciona, tomemos un único *PR*, por ejemplo Inglés. Durante el entrenamiento cada uno de los N idiomas de que se dispone $\{L_1, L_2, ..., L_N\}$, es *tokenizado* en secuencias de fonemas del inglés, las cuales son empleadas para entrenar los distintos LMs basados en n-gramas de fonemas $\{LM_1, LM_2, ..., LM_N\}$.

Ya en la fase de prueba, la muestra pasa igualmente por el *tokenizador* para obtener una secuencia de fonemas de longitud J, $\hat{Y} = \{p_1, p_2, ..., p_J\}$. Se tiene entonces que para cada idioma l, la probabilidad de una secuencia fonemas dado un modelo es:

$$logP(\hat{Y}|LM_l) = \sum_{j=1}^{J} logP_{LM_l}(p_j | p_{j-1} ... p_{j(n-1)}). \tag{7}$$

La ecuación 7, es interpretada como la entropía cruzada entre la secuencia de fonemas $\hat{Y}$ que representan la distribución empírica de la muestra de prueba, con el modelo fonético de n-gramas.

Es también muy empleada en este campo y con un significado análogo, la perplejidad [25]. Un bajo valor de perplejidad indica que un determinado n-grama de fonemas coteja mejor con la secuencia observada, en otras palabras, la secuencia de fonemas es más predecible.

Por otro lado, utilizar múltiples *tokenizadores* (PPRLM) permite tener las estadísticas de la muestra de prueba desde distintos puntos de vista[26], observar el mismo fenómeno con distintos lentes.

[12] *Phone Recognition* o *Parallel Phone Recognition followed by Phone n-gram Language Model.*

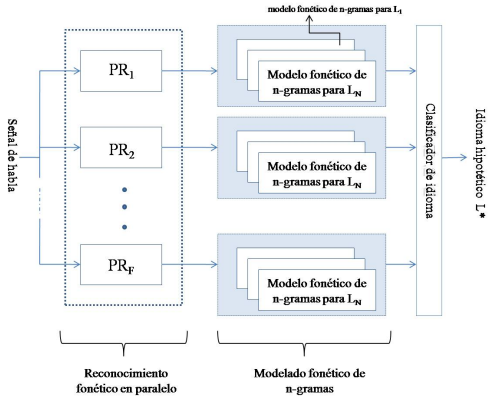

Fig. 3. Diagrama de un sistema PPR-LM para el reconocimiento de idioma.

De acuerdo a Fig3 para cada uno de los F reconocedores fonéticos, se entrenan N LMs, por lo que PPRLM puede ser visto como una fusión de varios subsistemas PRLM.

Dada una muestra de prueba, se generan FN puntuaciones de los FN modelos de n-gramas. ¿Cómo tomar una decisión entre tantas puntuaciones?

Es en este sentido que grandes esfuerzos se han realizado, buscando optimizar la forma de acoplar las salidas de los subsistemas [27] , las cuales son conocidas como ténicas de *back-end*. Una primera aproximación pudiera ser: sean $f = 1, 2, ..., F$ y $l = 1, 2, ..., N$, entonces

$$\log P(L_l|\mathcal{O}) = \sum_{f=1}^{F} \log \frac{P(\hat{Y}|LM_{f,l})}{\sum_{i=1}^{N} P(\hat{Y}|LM_{f,i})}, \tag{8}$$

donde $\hat{Y}_f$ resulta la secuencia de fonemas generada por el F-ésimo reconocedor de la observación $\mathcal{O}$ y $P(\hat{Y}|LM_{f,l})$ es la probabilidad para el l-ésimo idioma (7).

3.3 Modelado en el espacio de vectores

El enfoque análogo al conocido en la categorización de textos como saco-de-palabras [13], es otro intento exitoso en el reconocimiento de idiomas. Es sencillo establecer un vínculo entre las palabras en un saco-de-palabras y los *tokens* en un saco-de-sonidos, la diferencia estriba en que en el último caso son secuencias de fonos y no de palabras.

La esencia del modelado en el espacio vectorial es el mapeo de las estadísticas de las señales (sean de prueba o entrenamiento) a vectores de alta dimensionalidad. Obsérvese la arquitectura de un sistema que acopla al reconocimiento fonético con los modelos de los idiomas en el espacio de vectores (PPRVSM 4).

Sean F PRs, con inventario fonético $\nu = \{\nu_1, \nu_2, ..., \nu_F\}$ y n_f el número de fonemas en ν_F. La señal es *tokenizada* en F secuencias distintas, cada una de las cuales es representada en un vector que recoge las estadísticas de los fonemas.

Supongamos la situación en la que solo tenemos en cuenta unigramas y bigramas[14], tendremos pues un vector de $n_f + {n_f}^2$ elementos, que denotaremos V_f y que representa la secuencia del f-ésimo reconocedor fonético. Los F vectores se concatenan formando un vector-compuesto $V =$

[13] *bag-of-words*

[14] n-gramas donde $n = 1, 2$ respectivamente.

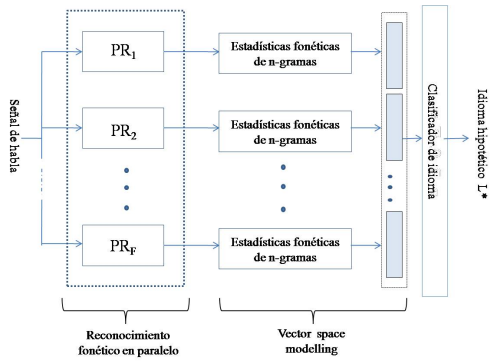

Fig. 4. Diagrama de un sistema PPR-VSM para el reconocimiento de idioma.

$[V_1^T, V_2^T, ..., V_F^T]^T$ de dimensión $B = \sum_f (n_f + {n_f}^2)$. Una vez que la señal es representada de esta forma, el reconocimiento de idioma puede ser visto como una clasificación de vectores, donde una primera aproximación pudiera ser medir la similaridad entre dos vectores-compuestos, uno derivado de la muestra de prueba y el otro derivado de la data de entrenamiento (uno contra todos). La similaridad entre dos vectores puede ser aproximada con el producto interno o la distancia coseno, sin embargo la estrategia del VSM se ha beneficiado muchísimo de la ventajas que brindan las máquinas de vectores soporte (SVM[15]) para clasificar vectores de alta dimensión [28].

Para cada idioma se entrena una SVM, usando los vectores del correspondiente idioma como positivos, y los correspondientes al resto de los idiomas como negativos. Dada la muestra de prueba, se tendrán tantas puntuaciones de salida como idiomas haya entrenados (N) y resulta muy útil organizar estas puntuaciones en forma de vector N-dimensional, ya que significaría una reducción importante de dimensionalidad. De manera similar a cómo se explicó en 3.2, aplicar técnicas de *back-end* al vector de puntuación representó una significativa mejora en el rendimiento de estos sistemas.

3.4 *Front-end* fonotáctico

Al grupo de pasos y métodos aplicados a la señal, previo a la clasificación propiamente dicha, se le llama *front-end*.

Tanto el modelado fonético con n-gramas (LM) como el modelado en el espacio de vectores (VSM) se apoyan en el mismo *front-end* (PR o PPR) para extraer las estadísticas, la diferencia radica en la forma de representar esos n-gramas. El estudio del saco-de-sonidos ha motivado investigaciones para incorporar mayor número de n-gramas evitando un fatalismo dimensional [29,30]. Mientras los n-gramas son típicamente estimados sobre la mejor transcripción, hay investigaciones que respaldan que se obtienen mejores resultados si los n-gramas se derivan de las redes de fonemas. La mejora se atribuye a la información adicional disponible que se halla en las redes [31].

Está bastante generalizado en la comunidad científica que los rasgos cepstrales (MFCC[16]) y los cepstrales Δ-desplazados (SDC[17]) son el punto de partida para el SLR, sin embargo se

[15] *Support Vector Machines*
[16] *Mel-frequency Cepstrum Coefficients*
[17] *Shifted Delta Cepstral Features*

trabaja en la selección de rasgos [32] a partir de aplicar métodos de maximización de la varianza y proyecciones a subespacios donde se mejora el desempeño de determinado clasificador [33].

Es también importante analizar la repercusión del ancho de la ventana temporal de la que se parte para extraer la información. Las representaciones de corto tiempo son ampliamente usadas en los sistemas actuales de reconocimiento de habla, sin embargo los experimentos señalan que se alcanzan mejores resultados cuando es incorporado al sistema información suprasegmental o de largo tiempo [34].

Tabla 2. Comparación de varios trabajos usando técnicas distintas para el reconocimiento de fonemas sobre TIMIT [35].

Año	**Sistema**	**Método**	**% Eficacia**
1989	Lee y Hon [36]	HMM	66.08
1991	Robinson y Fallside [37]	Recurrent Error Propagation Network	68.9
1992	Young [38]	HMM	61.07
1993	Lamel y Gauvain [39]	HMMs, trifonos	72.9
1994	Robinson, [40]	RNN	75.0
1998	Halberstadt, [41]	Clases alargadas	75.6
2003	Reynolds y Antoniou, [42]	MLP, clases alargadas	75.8
2006	Sha y Saul [43]	GMMs -SVMs	69.9
	Schwarz, Matejka e Cernocky,[44]	TRAPs + División del contexto temporal	78.52
2007	Deng, Yu y Acero [45]	Hidden Trajectory Models	75.17
	Rose y Momayyez [46]	TDNN, rasgos fonológicos HMM	72.2
	Scanlon, Ellis y Reilly, [47]	MLP/HMM	74.2
	ASAT, [48]	MLP/HMM	69.52
	Siniscalchi, Schwarz y Lee, [49]	TRAPs + División del contexto temporal + re-puntuación de la red	79.04
2008	Morris y Fosler-Lussier [50]	MLP/CRF	71.49
2009	Hifny y Renals, [51]	CRFs aumentadas	77.0
2010	Mohamed, Hinton, [52]	Máquinas de Boltzmann	77.3
2011	Mohamed, et al., [53]	Monophone Deep Belief Networks	79.3

Véase en Tabla 2 cómo los mejores resultados alcanzados en reconocimiento de fonemas se obtienen de sistemas basados en redes neuronales. En las redes neuronales el uso de una ventana contextual, que engloba varias ventanas, permite al sistema aprender, dentro de ciertos límites, los patrones temporales de las unidades del habla [49].

De manera general se apuesta porque un mayor número de reconocedores fonéticos en paralelo, n-gramas de mayor orden y PRs más eficaces, brinden rasgos fonotácticos más informativos lo que lleva a sistemas más eficaces.

3.5 Tendencias actuales

Las principales contribuciones en esta área contemplan a reconocedores fonéticos en paralelo y dedican grandes esfuerzos a hacer eficiente la incorporación de n-gramas de mayor orden [30]. También es notable la importación de métodos del reconocimiento del locutor, para disminuir la variabilidad de sesión y robustecer los rasgos [54], [29]. Respecto al back-end han habido

contribuciones en cuanto al uso de las SVM como clasificador de los modelos del idioma [55] [30] y con mucha fuerza el empleo del método VSM para modelar la estadística de los idiomas [21].

4 Enfoque acústico-fonético

Este paradigma parte de la hipótesis de que cada idioma tiene su propio repertorio fonético [11] y por tanto intenta modelar la distribución acústica-fonética de los idiomas, partiendo de rasgos acústicos.

Los primeros esfuerzos por aplicar dicho enfoque al reconocimiento datan de 1980. Fueron pioneros los rasgos LPC[18] [16], así como el contorno de pitch y formantes acústicos. Para la clasificación se han realizado estudios comparativos entre Cuantificación Vectorial (VQ[19]), HMM discretas, HMM continuas y GMM [56] apartir de rasgos MFCC, dándole especial crédito a las HMM continuas y a las GMM. Se verá en esta sección que hay grandes esfuerzos dirigidos a modificar el clasificador o a encontrar subespacios donde estos mejoren su desempeño.

4.1 Extracción de rasgos acústicos

Los rasgos MFCC [16] son efectivos en la mayoría de los temas relacionados con habla, ya que explotan principios estructurales del sistema auditivo, por lo tanto no sorprende que se desempeñen bien en el SLR.

Buscando robustez en cuanto a variaciones de sesión, se aplican técnicas de compensación (CMVN, RASTA, VTLN [57]) tradicionalmente precedidas de una eliminación de silencios.

Los MFCC son calculados sobre pequeños segmentos de la señal (10 ms) junto a sus derivadas de 1^{er} y 2^{do} orden (Δ y $\Delta\Delta$), y justamente es este parámetro del tiempo sobre el cuál se extrae la información acústica, el que más ha repercutido en su desempeño. Véase como los SDC [58] tienen mejores resultados debido a que abarcan intervalos mayores de señal, por tanto reflejan mejor la dinámica de la misma.

4.2 Modelado estadístico

El reconocimiento de idioma hablado y el reconocimiento de locutor, tienen muchas similitudes en términos de formulación técnica, metodología y medidas de evaluación. Un gran número de estrategias para enfrentar el problema han sido importadas con éxito de esa área, por ejemplo el modelo universal (UBM[20]) [57] a partir de las GMM.

Uno de los atributos atractivos de las GMM es su capacidad de modelar distribuciones arbitrarias de datos, a partir de aproximarse a las clases subyacentes con las componentes gaussianas individuales. En SLR, cada trama espectral es considerada como una muestra independiente, las GMM son empleadas para aproximar la distribución de dichas tramas para cada idioma, y han demostrado competitividad en las aplicaciones.

En SLR se entrena un GMM para cada idioma. En el paradigma GMM-UBM [59], donde UBM es un modelo que representa el universo de idiomas hablados, generalmente se comienza

[18] *Linear Predictive Coefficients*
[19] *Vector Quantitation*
[20] *Universal Background Model*

entrenando dicho modelo a partir de señales mutilenguaje. Luego se adapta dicho modelo genérico a cada idioma en particular usando MAP [60]. Esta metodología se ha convertido en referencia para el SLR.

Sin embargo, la dependecia de las GMM a la distribución de los rasgos acústicos, las hace marcadamente vulnerables a variaciones independientes del idioma, como son la variabilidad de canal y de locutor. Esto ha demandado la aplicación de técnicas de compensación de la variabilidad de sesión.

4.3 Modelado en el espacio de vectores

Así como en el enfoque fonotáctico ha resultado efectiva la representación vectorial de los datos, aquí también se busca representar en espacios vectoriales los rasgos espectrales. Las SVMs son comúnmente empleadas para problemas de clasificación binaria, estas le hacen corresponder al vector de entrada un valor escalar $f(x)$ de la forma:

$$f(\mathbf{x}) = \sum_{i=1}^{I} \alpha_i \kappa(\mathbf{x}, \mathbf{x_i}) + \beta, \tag{9}$$

de manera que $\mathbf{x_i}$ son los vectores soporte, I es el número de vectores soporte, α_i son pesos y β es un bias. Se exige que $\sum_{i=1}^{I} \alpha_i = 0, \alpha_i \neq 0$. La función $\kappa(\cdot)$ es el kernel, y sobre su elección y orden versan un gran número de investigaciones del estado del arte.

4.4 Tendencias actuales

En esta aproximación se modela cada idioma con Mezclas Gaussianas adaptadas de un modelo universal (procedimiento que se conoce como GMM-UBM-MAP), obteniéndose supervectores que luego entrenan al clasificador encargado de tomar la decisión. En la actualidad, los supervectores de media gaussianos adaptados, con un clasificador SVM, son la configuración más estándar. La aproximación acústica solo requiere datos de audio de los lenguajes a entrenar, no precisa información específica de idioma como transcripciones, lo que representa un incentivo importante para seguir mejorando el desempeño de los sistemas basados en ella.

Actualmente se dedican grandes esfuerzos investigativos a mejorar la modelación del idioma, influyendo por ejemplo en la adaptación MAP que se hace del UBM [61], a partir de demostrar que el factor de relevancia depende de los datos a adaptar.

La representación de la información acústica empleando supervectores, abre el camino para optimizar el espacio de los mismos [62], reduciendo dimensionalidad e incorporando la exitosa aproximación de varibilidad total [61,63].

Igualmente sobre el clasificador se han estudiado modificaciones: en [64] aplicando kernels no lineales a las SVMs se obtuvieron mejoras de hasta un 24 % de reducción del EER [21], una medida del error de verificación que da la exactitud del umbral para el cual los falsos rechazos y las falsas aceptaciones son iguales [65].

[21] *Equal Error Rate*

5 Enfoques alternativos

Una vez revisados los métodos y sistemas del estado del arte, resultó interesante el siguiente fenómeno comprobado con experimentos perceptuales:

El ser humano, luego de determinado tiempo expuesto a un idioma del cual no tiene ningún conocimiento lingüístico, es capaz de reconocerlo de otros igualmente desconocidos para él. O sea, se percata de cierta información contenida en el habla y no precisamente en las palabras, que le permiten identificar el idioma escuchado.

Entonces, ¿cuán necesarios son realmente los fonemas, o cualquier otra unidad lingüística definida, para modelar e identificar un idioma?

Nos planteamos como tarea explorar atributos presentes en el habla, dependientes del idioma sabiendo que los sistemas fonotácticos se mantienen con los mejores resultados en cuanto a eficacia en la identificación de idiomas y que la dependencia con señales fonéticamente transcritas y etiquetadas afecta tanto la obtención de buenos modelos fonéticos como la posibilidad de cotejar las condiciones de entrenamiento y prueba.

La anterior idea nos llevó a localizar nuestro dominio de acción en la tokenización con unidades acústicas como alternativa a los fonemas y a incorporar a esta, una nueva representación con información suprasegmental y de alto nivel. Igualmente es de interés modelar los idiomas con la técnica de minería de texto VSM.

Una representación que no esté amarrada a una definición fonética en particular, resultará más universal y conceptualmente más fácil de adoptar (recordar ambigüedad de la definición de fonema). Por su parte estas unidades acústicas son altamente deseables para una caracterización universal del lenguaje, especialmente para aquellos idiomas poco comunes o que no cuenten con una ortografía o diccionarios fonéticos bien definidos [18].

Para evaluar preliminarmente estas ideas, llevamos a cabo algunos experimentos.

5.1 Experimentación

En el diseño experimental se buscó trabajar con un *front-end* puramente acústico, que brindara una representación útil para extraer información fonotáctica en futuras experimentaciones.

La señal de habla es dividida en tramas comúnmente de 25ms, con un desplazamiento de ventana de 10ms, y se asume que la señal es estacionaria en dichas tramas. Una vez aquí, se extrae un grupo de parámetros que describen cada trama. El objetivo es reducir dimensionalidad para tener un mejor desempeño de los clasificadores y minimizando la influencia del canal, variabilidad intra-locutor, etc. En la actualidad, la técnica de extracción de rasgos más popular pudiera decirse que es MFCC. Estos vectores de rasgos pueden ser vistos como puntos en un espacio N-dimensional, donde N es la dimensión de los rasgos (Fig.5).

Como parte de la información que portan los rasgos, está la relacionada al estado de nuestros órganos articulatorios. Como el movimiento de los órganos de nuestro aparato fonador es lento, se asume, sin perder generalidad, que rasgos consecutivos en el dominio del tiempo, serán cercanos en el dominio cepstral también.

Si hubiese más tramas similares (como pasa en una vocal), más puntos estarían cercanos. Por el contrario, puntos que descansen en las transiciones entre fonemas estarán separados. Una secuencia de esos puntos, representa una trayectoria. La trayectoria es un resultado del proceso generador del habla, imaginemos pues un punto en movimiento, con velocidad variable, en el espacio N-dimensional de rasgos.

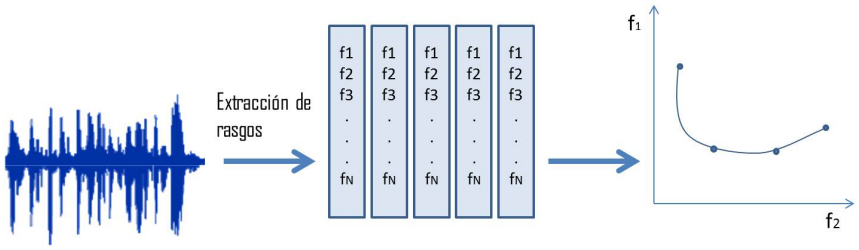

Fig. 5. ¿Cómo ver la parametrización? Habla, vector de rasgos, punto móvil en espacio N-dimensional.

La velocidad es mayor en los puntos de transición entre zonas no estacionarias, y menor en zonas estacionarias. Para nosotros, en la tarea de reconocimiento, una frase, un palabra o un fonema, es una parte de esa trayectoria, y es justamente lo que se busca modelar lo más precisamente posible. La velocidad de movimiento del punto, es también importante, ya que porta información relevante sobre la duración de los fonemas.

Seguimos 3 etapas fundamentales en la experimentación:

- segmentar la señal en intervalos espectralmente estables,
- agrupar clases acústicas,
- modelar el idioma,
- evaluación.

Segmentación en eventos acústicos cepstralmente estables Basándonos en las ideas anteriores, decidimos aplicarle al conjunto de rasgos MFCC extraídos, una función de variabilidad espectral (SVF[22]) que sería la encargada de determinar, apartir del valor definido como umbral, intervalos de "estabilidad"del espectrograma.

Básicamente SVF mapea los rasgos MFCC al espacio de la diferencia de vectores contiguos, y luego delimita segmentos donde los vectores diferencia son menores que la diferencia media de los datos. De acuerdo a lo anterior dichos vectores serán "cercanos". Esta diferencia es caculada en primera instancia con la distancia euclideana.

Dado que el procedimiento de segmentación arrojó intervalos supuestamente estables, tiene sentido resumir el contenido espectral del segmento en un vector, por tanto se conserva solo el vector promedio de cada intervalo. Y en estos términos finalmente queda representada la señal.

Agrupar clases acústicas El primer paso en la modelación fue entrenar el UBM con un conjunto de señales de la base CSLU [66] de inglés y español, puesto que la tarea que nos propusimos fue una clasificación binaria en español o inglés. A este modelo le llamaremos *Espanglish*. Modelamos con gaussianas la distribución de los rasgos con diferente número de mezclas (15, 64 y 128) buscando la configuración óptima.

Modelar el idioma Para modelar cada idioma se propusieron 2 estrategias iniciales.

Estrategia 1: Usando 10 señales por idioma, se adaptaron las medias del UBM a cada una (UBM-MAP), y buscando una representación genérica del idioma se promediaron dichos valores para constituir el modelo de cada idioma.

Estrategia 2: Tomar 10 señales de cada idioma, concatenarlas y adaptar (UBM-MAP) el modelo *Espanglish* a su distribución. Las medias adaptadas serán el modelo para cada idioma.

Evaluación La evaluación se hizo con 3 bases de datos:

- TIMIT [35] remuestreada a 8kHz, 100 frases en inglés.

[22] *Spectral Variation Function*

- Sala [67], 100 frases en español.
- Ahumada [68], 300 frases en español.

La misma consistió en calcular la probabilidad de los rasgos de la muestra de prueba, dado ambos modelos adaptados (español e inglés). El idioma asociado al modelo de más probable fue seleccionado como el idioma de la muestra de prueba en correspondencia con 3. Este procedimiento se siguió para los modelos construídos con ambas estrategias.

Los resultados son reportados en % de señales con el idioma correctamente identificado. Para denotar los modelos en la tabla se empleó la siguiente metodología: $\#Mezclasgaussianas\_Estrategia$.

Tabla 3. Resultados de la clasificación binaria de idiomas (Español o Inglés).

Modelo	TIMIT	Sala	Ahumada
15_1	95	96	
64_1	94	97	98
128_1	92	94	
15_2	96	55	
64_2	98	87	
128_2	100	45	

No resultaría muy sano comparar el desempeño de los modelos frente a uno u otro idioma, pues las señales no son de igual duración. Sin embargo se pudiera especular en cuanto a que el número de gaussianas óptimo debiera estar cercano al cardinal del repertorio fonético de cada idioma, ya que los fonemas son una representación de los sonidos que se usan.

Estos resultados, si bien no cumplen los requisitos de las actuales competencias internacionales del ámbito, constituyen los primeros pasos de nuestro grupo en el desarrollo de algoritmos que demandan muy bajos recursos.

También con ello se delimitan nuevas tareas para mejorar el rendimiento y ampliar el campo de acción, como por ejemplo la adición de un tercer idioma y la introducción de técnicas de pesaje, buscando realzar la gaussianas no tan frecuentes pero con probada información valiosa para el SLR.

Igualmente queda abierto el campo de mayor interés para nosotros, que es el trabajo en el *front-end*, definiendo una nueva SVF que incorpore las derivadas de los rasgos, o sea que incluya información dinámica.

5.2 Trabajos futuros

Nuestra idea es convertir la señal de habla en una secuencia de unidades acústicas, especificadas en un inventario acústico universal, independiente del idioma, que recoja información prosódica y articulatoria y las combine eficientemente. Con estos atributos corresponderá entrenar los modelos del idioma. El *tokenizador* en nuestra propuesta estará formado por un Segmentador de Unidades Acústicas (AUS[23]) y por un Agrupador de Unidades Acústicas (AUC[24]), que trabajarán de forma no supervisada y sin demandar más información que la contenida en la señal de audio.

[23] *Acoustic Units Segmenter*
[24] *Acoustic Units Clustering*

Existe un grupo de trabajos en esta dirección, buscando independizarse del reconocimiento fonético, sin embargo los intentos con mejores resultados parten de datos fonéticamente etiquetados. Han habido esfuerzos por prescindir de las transcripciones, enfocando el problema en una segmentación adecuada para luego modelar.

Los principales trabajos precedentes relacionados pueden ser clasificados en los siguientes grupos:

- Aproximan una segmentación fonética basándose en variaciones del espectrograma.
  Este trabajo busca en el espectrograma intervalos de señal donde los rasgos MFCC varíen poco. Esos rasgos seleccionados son modelados con GMM, y las clases generadas durante el proceso de entrenamiento son los *tokens* que permitirán construir los LM, lo que constituye una solución totalmente independiente de las etiquetas.
  [69] AERLM 31.5 %, PPRLM 20 %
- Representan el espacio acústico de todos los lenguajes con GMM.
  En ambos trabajos se modela la distribución de los rasgos con GMM, y las clases acústicas son empleadas como *tokens*. Incorpora la filosofía del PPRLM a partir de experimentar con múltiples *tokenizadores* gaussianos en paralelo. Las diferencias radican en la forma de construir el LM (n-gramas) y en la de clasificar.
  [17,70] GMM-tokenizer 26.7 %, PPRLM 22 %
- Segmentan y etiquetan la señal basados en características prosódicas (dinámica de la frecuencia fdtal y la energía).
  Es llamativo el conjunto pequeño de clases con que representan las señales. El objetivo de este enfoque es complementar a los sistemas convencionales.
  [71,72] SC 24.2 %, PRLM 21.3 %
- Retoman el concepto de ASM propuesto por Lee (1988). Los dos primeros trabajos conforman un *superset* de fonemas de muchos idiomas buscando una representación universal, pero partiendo de gran cantidad de datos etiquetados. Por su parte el último trabajo también se apoya en las transcripciones, pero para mapearlas a atributos articulatorios.
  [73] ASM 19.2 %, PPRLM 22 %
  [21] UPRVSM 24.44 %, PPRVSM 21.66 %
  [18] UAR 6.4 %, PPRLM 6.9 %

Véase al lado de cada referencia, el valor del EER obtenido con el modelo que se propone y luego el obtenido con PPRLM considerado como línea base en todos los casos. Vale aclarar que los dos únicos modelos que aventajan a PPRLM fueron obtenidos con una inicialización que partía de etiquetas fonéticas o articulatorias, o sea que del todo no prescinden de ellas, el reto sigue siendo alcanzar esos valores de EER sin emplearlas en lo absoluto.

Referencias bibliográficas

1. Zhao, J., Shu, H., Zhang, L., Wang, X., Gong, Q., Li, P.: Cortical competition during language discrimination. NeuroImage **43**(3) (2008) 624–633
2. A. Waibel, P. Geutner, L.M.T.T.S.M.W.: Multilinguality in speech and spoken language systems. In: Proceedings of the IEEE, Vol. 88, No. 8., pp. 1297-1313. (2000)
3. Ma, B., Guan, C., Li, H., Lee, C.H.: Multilingual speech recognition with language identification. In: INTERSPEECH. (2002)
4. Comrie, B., ed.: The world's major languages. Oxford University Press, New York (1990)
5. : The Cambridge Factfinder. 3/E. Cambridge University Press (1998)

6. Translators, W.B., of Linguistics, S.I.: Ethnologue. Number v. 11. Summer Institute of Linguistics (2009)
7. Lui, M., Baldwin, T.: langid.py: An off-the-shelf language identification tool. In: ACL (System Demonstrations). (2012) 25–30
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
9. Muthusamy, Y., Jain, N., Cole, R.A.: Perceptual benchmarks for automatic language identification. In: In International Conference on Speech and Signal Processing. (1994) 333–336
10. Leeuwen, D.A.V., Boer, M.D., Orr, R.: A human benchmark for the nist language recognition evaluation 2005. In: In Proc. Speaker and Language Odyssey, Stellenbosch, South Afrika, IEEE (2008)
11. Schultz, T., Kirchhoff, K.: Multilingual Speech Processing. Elsevier Science (2006)
12. Ashby, M., Maidment, J.: Introducing Phonetic Science. Cambridge Introductions to Language and Linguistics. Cambridge University Press (2005)
13. Benesty, J., Sondhi, M.M., Huang, Y., eds.: Springer Handbook of Speech Processing. Springer, Berlin (2008)
14. Mariani, J.: Spoken Language Processing. Wiley (2008)
15. Oneisys Núñez Cuadra, J.R.C.d.L.: Métodos de reconocimiento automático del idioma. Technical report, Centro de Aplicaciones de Tecnologías de Avanzada (Agosto 2009)
16. Oneisys Núñez Cuadra, J.R.C.d.L.: Métodos de extracción de rasgos para la identificación del idioma: estado del arte. Technical report, Centro de Aplicaciones de Tecnología de Avanzada (Junio 2010)
17. Torres-Carrasquillo, P.A., Reynolds, D.A., Deller, J.R.: Language identification using gaussian mixture model tokenization. In: ICASSP. (2002) 757–760
18. Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.H.: Universal attribute characterization of spoken languages for automatic spoken language recognition. Computer Speech & Language **27**(1) (2013) 209–227
19. Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.H.: Exploring universal attribute characterization of spoken languages for spoken language recognition. In: INTERSPEECH. (2009) 168–171
20. Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.H.: Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition. In: INTERSPEECH. (2010) 2718–2721
21. Li, H., Ma, B., Lee, C.H.: A vector space modeling approach to spoken language identification. IEEE Transactions on Audio, Speech & Language Processing **15**(1) (2007) 271–284
22. Schwarz, P.: Phoneme Recognition based on Long Temporal Context. PhD thesis, Brno University of Technology, Czech Republic (2008) Supervisor - JAN CERNOCKY.
23. Hermansky, H., Sharma, S.: Temporal patterns (traps) in asr of noisy speech. In: in Proc. ICASSP. (1999) 289–292
24. Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK (2006)
25. Jurafsky, D., Martin, J.H.: Speech and language processing. Prentice Hall, Upper Saddle River, NJ (2000)
26. Zissman, M.A.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing **4**(1) (1996) 31
27. Peñagarikano, M., Varona, A., Díez, M., Rodríguez-Fuentes, L.J., Bordel, G.: Study of different backends in a state-of-the-art language recognition system. In: INTERSPEECH. (2012)
28. Varona, A., Peñagarikano, M., Rodríguez, L.J., Bordel, G.: On the use of lattices of time-synchronous cross-decoder phone co-occurrences in a svm-phonotactic language recognition system. In: INTERSPEECH. (2011) 2901–2904
29. Soufifar, M., Cumani, S., Burget, L., Cernocký, H.: Discriminative classifiers for phonotactic language recognition with ivectors. In: ICASSP. (2012) 4853–4856
30. Peñagarikano, M., Varona, A., Rodríguez, L.J., Bordel, G.: Dimensionality reduction for using high-order n-grams in svm-based phonotactic language recognition. In: INTERSPEECH. (2011) 853–856
31. Gauvain, J.L., Messaoudi, A., Schwenk, H.: Language Recognition Using Phone Lattices. In: ICSLP, Jeju Island (2004) 1283–1286
32. Mikolov, T., Plchot, O., Glembek, O., Matejka, P., Burget, L., Cernocky, J.: Pca-based feature extraction for phonotactic language recognition. In: Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop. (2010) 251–255
33. Shih, Y.C., Lee, H.S., Wang, H.M., Jeng, S.K.: Subspace-based feature representation and learning for language recognition. In: INTERSPEECH. (2012)
34. Lopes, C.A.C.: Classes foneticas alargadas no reconhecimento automatico de fones. PhD thesis, Universidade de Coimbra, Portugal (2011) Supervisor - Doutor Fernando Manuel dos Santos Perdigão.
35. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993)

36. Lee, K.F., Hon, H.W.: Speaker-independent phone recognition using hidden Markov models. IEEE Transactions on Acoustics, Speech and Signal Processing **37**(11) (November 1989) 1641–1648
37. Robinson, T., Fallside, F.: A recurrent error propagation network speech recognition system. Computer Speech & Language **5**(3) (1991) 259–274
38. Young, S.J.: The general use of tying in phoneme-based hmm speech recognisers. In: Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1. ICASSP'92, Washington, DC, USA, IEEE Computer Society (1992) 569–572
39. Lamel, L.F., Gauvain, J.: High performance speaker-independent phone recognition using cdhmm. In: In Proc. Eurospeech. (1993) 121–124
40. Robinson, T.: An application of recurrent nets to phone probability estimation. IEEE Transactions on Neural Networks **5** (1994) 298–305
41. Halberstadt, A.K., Glass, J.R.: Heterogeneous measurements and multiple classifiers for speech recognition. In: ICSLP. (1998)
42. Reynolds, T.J., Antoniou, C.A.: Experiments in speech recognition using a modular mlp architecture for acoustic modelling. Inf. Sci. Inf. Comput. Sci. **156**(1-2) (November 2003) 39–54
43. Sha, F., Saul, L.K.: Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition. In: Proc. ICASSP-2006. 265–268
44. Schwarz, P., Matějka, P., Černocký, J.: Hierarchical structures of neural networks for phoneme recognition. In: Proceedings of ICASSP 2006. (2006) 325–328
45. Deng, L., Yu, D., Acero, A.: A generative modeling framework for structured hidden speech dynamics. In: In Proceedings of NIPS Workshop on Advances in Structured Learning for Text and Speech Processing. (2005)
46. Rose, R., et al.: Integration of multiple feature sets for reducing ambiguity in asr (2007)
47. Scanlon, P., Ellis, D.P.W., Reilly, R.B.: Using broad phonetic group experts for improved speech recognition. IEEE Transactions on Audio, Speech & Language Processing **15**(3) (2007) 803–812
48. Bromberg, I., Qian, Q., Hou, J., Li, J., Ma, C., Matthews, B., Moreno-Daniel, A., Morris, J., Siniscalchi, S.M., Tsao, Y., Wang, Y.: Detection-based asr in the automatic speech attribute transcription project. In: INTERSPEECH. (2007) 1829–1832
49. Siniscalchi, M.S., Schwarz, P., Lee, C.H.: High-accuracy phone recognition by combining high performance lattice generation and knowledge based rescoring. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), IEEE Signal Processing Society (2007) 869–872
50. Morris, J., Fosler-Lussier, E.: Combining phonetic attributes using conditional random fields. In: INTERSPEECH. (2006)
51. Hifny, Y., Renals, S.: Speech recognition using augmented conditional random fields. IEEE Transactions on Audio, Speech & Language Processing **17**(2) (2009) 354–365
52. rahman Mohamed, A., Hinton, G.E.: Phone recognition using restricted boltzmann machines. In: ICASSP. (2010) 4354–4357
53. rahman Mohamed, A., Dahl, G.E., Hinton, G.E.: Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech & Language Processing **20**(1) (2012) 14–22
54. Soufifar, M., Kockmann, M., Burget, L., Plchot, O., Glembek, O., Svendsen, T.: ivector approach to phonotactic language recognition. In: INTERSPEECH. (2011) 2913–2916
55. Boril, H., Sangwan, A., Hansen, J.H.L.: Arabic dialect identification - 'is the secret in the silence?' and other observations. In: INTERSPEECH. (2012)
56. Sugiyama, M.: Automatic language recognition using acoustic features. In: Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference. ICASSP '91, Washington, DC, USA, IEEE Computer Society (1991) 813–816
57. Ana Montalvo Bereau, J.R.C.d.L.: Métodos para reducir la variabilidad de sesión en el reconocimiento del locutor. Technical report, Centro de Aplicaciones de Tecnologías de Avanzada (Agosto 2012)
58. Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Jr., J.R.D.: Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In: INTERSPEECH. (2002)
59. José Ramón Calvo, Rafael Fernández, G.H.: Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor. Technical report, Centro de Aplicaciones de Tecnologías de Avanzada (Febrero 2008)
60. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing **2**(2) (1994) 291–298
61. You, C., Li, H., Ma, B., Lee, K.A.: Effect of relevance factor of maximum a posteriori adaptation for gmm-svm in speaker and language recognition. In: INTERSPEECH, ISCA (2012)

62. Plchot, O., Karafiát, M., Brummer, N., Glembek, O., Matějka, P., de, E.V., Černocký, J.: Speaker vectors from subspace gaussian mixture model as complementary features for language identification. In: Proceedings of Odyssey 2012, The Speaker and Language Recognition Workshop, International Speech Communication Association (2012) 330–333
63. Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D.A., Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: INTERSPEECH. (2011) 857–860
64. Yaman, S., Pelecanos, J.W., Omar, M.K.: On the use of non-linear polynomial kernel svms in language recognition. In: INTERSPEECH. (2012)
65. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech Commun. **52**(1) (January 2010) 12–40
66. Muthusamy, Y.K., Cole, R.A., Oshika, B.T.: The ogi multi-language telephone speech corpus. (1992) 895–898
67. Moreno, A., Comeyne, R., Haslam, K., van den Heuvel, H., Höge, H., Horbach, S., Micca, G.: Sala: Speechdat across latin america. results of the first phase. In: LREC. (2000)
68. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V.: Ahumada: A large speech corpus in spanish for speaker characterization and identification. Speech Communication **31**(2-3) (2000) 255–264
69. Danilo Spada, I. Lopez, D.T.J.G.: Acoustic event recognition for low cost language identification. In: V Jornadas en en Tecnologías del Habla 2007. UAM. (2007) 25–28
70. Hanani, A., Carey, M.J., Russell, M.J.: Improved language recognition using mixture components statistics. In: INTERSPEECH. (2010) 741–744
71. Adami, A.G., Hermansky, H.: Segmentation of speech for speaker and language recognition. In: INTERSPEECH. (2003)
72. Ng, R.W.M., Leung, C.C., Lee, T., Ma, B., Li, H.: Prosodic attribute model for spoken language identification. In: ICASSP. (2010) 5022–5025
73. Ma, B., Li, H., Lee, C.H.: An acoustic segment modeling approach to automatic language identification. In: INTERSPEECH. (2005) 2829–2832

RT_057, octubre 2013
Aprobado por el Consejo Científico CENATAV

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. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba