RIELAC, Vol. 40 3/2019 p. 51-71 Septiembre – Diciembre ISSN: 1815-5928

RIELAC
Revista de Ingeniería Electrónica, Automática y Comunicaciones

Evaluación de Rasgos Acústicos para el Reconocimiento Automático del Habla en Escenarios Ruidosos usando Kaldi

José Manuel Ramírez Sánchez, Ana Rosa Montalvo Bereau, José Ramón Calvo de Lara

RESUMEN / *ABSTRACT*

La presente investigación evaluará el impacto de los Coeficientes Cepstrales en la Frecuencia Mel (MFCC) y los coeficientes Predictores Perceptuales Lineales (PLP), en la tasa de errores de reconocimiento de palabras (WER) de sistemas dedicados al Reconocimiento Automático del Habla (RAH). La experimentación se realizará con señales de voz en idioma español, en escenarios con niveles de ruido desconocidos y utilizando la herramienta del estado del arte Kaldi. El artículo concluye aportando evidencias a favor de los MFCC como rasgo acústico más robusto ante la tarea del RAH en escenarios ruidosos con respecto a los PLP; haciendo notar que ambos rasgos se comportar de manera similar en escenarios poco ruidosos y el impacto de los PLP en la reducción de los tiempos empleados por los sistemas dedicados al RAH.

Palabras claves: Reconocimiento Automático del Habla, Rasgos Acústicos, Kaldi.

Evaluation of Acoustic Features for the Automatic Speech Recognition in Noise Scenarios using Kaldi

The present investigation will evaluate the impact of Mel Frequency Cepstral Coefficients (MFCC) and the Perceptual Linear Predictors (PLP) coefficients, in the word error rate (WER) of systems dedicated to Automatic Speech Recognition (ASR). The experimentation will be done with voice signals in Spanish language, in scenarios with unknown noise levels and using the Kaldi state of the art tool. The article concludes by providing evidence in favor of the MFCC as acoustic feature more robust in the task of ASR in noisy scenarios with respect to the PLP; also both features behave similarly in low noise scenarios and the impact of PLP in reducing the time spent by systems dedicated to ASR.

Key words: Automatic Speech Recognition, Acoustic Features, Kaldi.

1. -INTRODUCCIÓN

El Reconocimiento Automático del Habla (RAH) es un área de investigación que pretende resolver la tarea de reconocer y entender el habla contenida en una señal acústica dado cualquier locutor, ambiente o idioma. Un problema fundamental en el área del RAH consiste en obtener una codificación compacta de la forma de onda de la señal de audio que contiene habla (señal de habla), que minimice las pérdidas de información. El resultado de esta codificación se conoce como rasgos acústicos y la comunidad científica, en el área del RAH, reconoce como estado del arte a los Coeficientes Cepstrales en escala de Frecuencias Mel (MFCC) [1] y los coeficientes Predictores Lineales Perceptivos (PLP) [2].

La efectividad de cualquier sistema dedicado al reconocimiento puede medirse en término de errores cometidos en relación al total de elementos a reconocer. En el caso del RAH esta medida se conoce como la tasa de errores de reconocimiento de palabras (WER), definida como:

Recibido: 15/3/2019 Aceptado: 14/6/2019

$$WER = 100 \cdot \frac{(INS + SUB + DEL)}{N}, \quad (1)$$

donde N es el número de palabras presentes en la locución a transcribir; INS, SUB y DEL son el número de errores cometidos por inserción, sustitución y omisión de palabras en la transcripción hecha por el sistema de RAH. Los actuales sistemas de RAH ofrecen WER alrededor de un 50% para las tareas más complejas y cerca de un 1% para tareas de transcripción de dígitos y letras, condiciones controladas.

Uno de los problemas por resolver del campo del RAH está determinado por la ausencia de consenso en la comunidad científica sobre cuál rasgo acústico, MFCC o los coeficientes PLP, ofrecen mayor efectividad en el RAH. El valor teórico de la investigación aquí presentada consiste en que aporta evidencias sobre la robustez de los MFCC como rasgo acústico empleado en el RAH en escenarios ruidosos en comparación con los PLP. La sección 2 describirá los procedimientos para la obtención de los rasgos acústicos utilizados y las transformaciones lineales empleadas sobre estos con el propósito de obtener mejores representaciones de la señal de habla. La sección 3 presentará la herramienta y los sistemas de RAH utilizados en la tarea de RAH. En la sección 4 se abordará el diseño de los experimentos realizados y describirá las características de las bases de datos usadas. La sección 5 se dedicará a discutir los resultados y finalmente en la sección 6 se ofrecen las conclusiones de la investigación.

2.- RASGOS ACÚSTICOS DEL HABLA

La extracción de rasgos es un aspecto fundamental de cualquier proceso de reconocimiento de patrones pues define qué características serán tenidas en cuenta para describir al objeto del reconocimiento. En el caso del RAH el objeto de reconocimiento es el habla y las características a tener en cuenta están relacionadas con el proceso de percepción humano del habla. Estas características no pueden ser extraídas directamente, por la complejidad natural de la forma de la onda del habla y por la gran variabilidad estadística de los parámetros físicos que describen dicha forma de onda. Estas complejidades pueden ser atenuadas mediante una adecuada representación espectral de la forma de onda, en el caso del RAH la representación espectral de término corto es de las más usadas, y una serie de transformaciones sobre dicha representación.

La extracción de rasgos acústicos consiste en una transformación de un espacio de alta dimensionalidad (las tramas de la señal de audio) a otro menor (los rasgos acústicos) [3]. Esto reduce el volumen de cálculo y el número de vectores necesarios para una caracterización robusta de la señal de habla. El resultado de esta transformación es conocido como rasgo acústico y no es más que un vector real de dimensión finita que representa de forma apropiada una trama de la señal de habla. Todo rasgo acústico debe permanecer invariante ante cambios de locutor, ambientes y ruido de fondo; debido a que las condiciones naturales de comunicación implican estos cambios. A esta característica fundamental de los rasgos acústicos se le conoce como robustez y es un elemento vital para el RAH en escenarios ruidosos.

Durante la extracción de los rasgos acústicos se busca obtener una parametrización espectral más decorrelacionada y de menor dimensión que las muestras de la señal de habla. Los métodos más usados para obtener una parametrización espectral de cada muestra son: el Banco de Filtros en la Escala de Frecuencias Mel y La Codificación Predictiva Lineal (LPC). Una vez definido el tipo de representación del objeto de reconocimiento y el conjunto de transformaciones sobre dicha representación pueden ser extraídas las características a reconocer. La comunidad científica, en el área del RAH, reconoce como estado del arte a los rasgos acústicos: Coeficientes Cepstrales en escala de Frecuencias Mel (MFCC) [1] y a los coeficientes Predictores Lineales Perceptivos (PLP) [2]. En el caso de los MFCC las características a reconocer son los aspectos acústicos de la percepción humana relevantes en la representación de corta duración del espectro de la señal de habla [1], mientras que los coeficientes PLP se concentran en los aspectos psicofísicos contenidos en la representación de corta duración del espectro de la señal de habla [2].

A partir de estos rasgos, se han desarrollado diferentes transformaciones, cuyo fin es mejorar las características discriminatorias y reducir la dimensión de los rasgos acústicos. Algunas de estas transformaciones son: Normalización Cepstral de Media y Varianza (CNMV) [4], Análisis Discriminativo Lineal (LDA) [5], Transformación Lineal de Máxima Verosimilitud (MLLT) [6], Espectro Relativo (RASTA) [7] o Regresión Lineal de Máxima Verosimilitud (MLLR) [8].

2.1.- MFCC

En 1980 Davis y Mermesltein en [1] propusieron un tipo de representación cepstral de la señal de habla que ha logrado mantenerse hasta la actualidad, como estado del arte de los rasgos acústicos empleados en el RAH, los MFCC. La idea fundamental de los MFCC consiste en procesar la señal de habla en múltiples bandas de frecuencias mediante un arreglo de filtros con pesos y anchos de banda diferentes (banco de filtros), de tal manera que se emule la percepción humana selectiva sobre ciertas bandas de la señal de habla [9]. Este procedimiento se logra mapeando a una escala no lineal de frecuencias conocida como escala de frecuencia de Mel según:

$$m = 1127 \cdot log\left(1 + \frac{f}{700}\right), \quad (2)$$

Un elemento crucial en los MFCC es el análisis cepstral y el concepto de cepstrum sobre el que se basa la representación. En el siguiente acápite será introducido el procedimiento y las características del rasgo acústico MFCC.

Procedimiento de obtención del rasgo acústico.
El procedimiento para la obtención de los MFCC puede ser descrito a grandes rasgos, como muestra la figura 1, como una secuencia de transformaciones sobre las muestras de la señal de habla. La mayoría de las implementaciones de los MFCC comienzan con la aplicación de una Ventana de Hamming de longitud 20 ms sobre el segmento de habla dado, y luego se realiza un análisis de Fourier de corta duración. Resultando en una Transformada Discreta de Fourrier (DFT) para la trama k-ésima.

Figura 1
PROCEDIMIENTO MFCC (TOMADA DE [1]).

Luego, los valores de DFT se agrupan en bandas críticas y se ponderan mediante funciones de ponderación triangulares. Los anchos de banda de dichas funciones son constantes para frecuencias centrales inferiores a 1 kHz y luego aumentan exponencialmente hasta la mitad de la frecuencia de muestreo. Una relación práctica comúnmente empleada [1,9] es utilizar 8 filtros por octava, resultando en la aplicación de 24 filtros para señales telefónicas y 40 filtros para señales microfónicas. El espectro de Mel de la k-ésima trama definido para r = 1, 2, ..., R como:

$$MF_k[r] = \frac{1}{A_r} \sum_{m=L_r}^{U_r} |V_r[m] X_k[m]|^2, \tag{3}$$

donde Vr[m] es una función de pesos para el filtro de orden r desde el índice inicial Lr hasta el índice final de la DFT. Vr[m] permite al filtro establecer un peso adecuado para cada subanda, de esta manera conocido el poder discriminativo asociado a dicha subanda puede atenuarla o amplificarla. El término Ar[m] definido como:

$$A_r = \frac{1}{r} \sum_{m=L_r}^{U_r} |V_r[m]|^2, \tag{4}$$

es un factor de normalización para el Filtro Mel de orden r. De cada trama es posible extraer los MFCC aplicando una transformada decorreladora, la transformada de coseno (DTC), al logaritmo de la magnitud de salida del Banco de Filtros Mel:

$$mfcc[n] = \frac{1}{R} \sum_{r=1}^{R} log(MF_k[r]) cos\left[\frac{2\pi}{R}\left(r + \frac{1}{2}\right) n\right], \tag{5}$$

El coeficiente cepstral de orden cero tiene una connotación importante pues representa el promedio de la energía del espectro, aunque puede ser despreciado con la intención de solo procesar información acústica contenida en la señal de habla. El número de coeficientes Nmfcc obtenidos de (5) se toma menor que el número de Filtros Mel; típicamente Nmfcc = 13 con R = 24 para señales telefónicas mientras que Nmfcc = 13 con R = 40 para señales microfónicas [10]. El espectro reconstruido a partir de los MFCC presenta ventajas como picos en los formantes principales o un suavizado sobre las bandas de altas frecuencias.

2.2.- PLP

El uso de predictores lineales como una forma de representación de la señal de habla que derive en un tipo de rasgo acústico parece natural y obvio, debido a la posibilidad de modelar el valor de una muestra de la señal de habla como combinación lineal de los valores de muestras precedentes a ella. En este acápite serán discutidos los coeficientes PLP como un rasgo acústico ampliamente reconocido y su fundamento teórico, la Codificación Predictiva Lineal (LPC). Será descrita la metodología Espectro Relativo (RASTA) aplicada sobre los coeficientes PLP en los RASTA-PLP como mecanismo para robustecer a este rasgo acústico ante los cambios de la respuesta de frecuencia que provoca el canal de comunicación y frente a efectos indeseados del entorno tales como ruido [7].

Procedimiento de obtención del rasgo acústico
La Predicción Lineal Perceptiva (PLP) es un método de estimación espectral de la señal de habla que en comparación con la LPC convencional es más consecuente con la percepción humana [2]. Este método se vale de tres conceptos psicofísicos de la audición humana: la resolución espectral en las bandas críticas, las curvas de distribución uniforme de intensidad y la ley de intensidad de potencia de ruido [2]. Este enfoque permite a los coeficientes PLP ser más efectivos en la supresión de los detalles espectrales dependientes del locutor [2], elevando la generalización de las estimaciones espectrales sobre las formantes de los fonemas y por tanto la efectividad en las tareas de reconocimiento con independencia del locutor. El empleo de los tres conceptos psicofísicos en el algoritmo de la PLP permite emular los siguientes rasgos psicofísicos de la percepción humana del habla:

- La resolución de frecuencia es lineal hasta 800 Hz y logarítmica por encima de los 800 Hz.
- Las frecuencias bajas enmascaran a las más altas (filtros asimétricos de banda crítica).
- La percepción del habla es más sensible a frecuencias medias (curvas de distribución uniforme de intensidad).
- La intensidad de ruido v[n] en la señal de habla observada y [n] es proporcional a la raíz cúbica de la intensidad de la señal de habla x[n].
- Integración espectral de más de 1 bark como hipótesis en la percepción del habla.

Estos conceptos permiten a las representaciones de la señal de habla basadas en la PLP codificar una información obviada por el resto de los métodos de representación espectral de la señal de habla. La figura 2 describe el algoritmo para obtener los coeficientes PLP a partir de las muestras de la señal de habla luego de la cuantificación, según [2].

La PLP comienza con la aplicación de una Ventana de Hamming de longitud 20 ms sobre el segmento de habla. La DFT transforma al dominio espectral cada ventana del segmento de habla, generalmente es usada la Transformada Rápida de Fourrier (FFT). La componente real e imaginaria del espectro de corta duración de la señal son elevadas al cuadrado y sumadas para obtener el espectro de potencia de corta duración:

$$P[w] = R(x[n])^2 + I(x[n])^2, \tag{6}$$

El espectro de potencia es mapeado del dominio espectral ω al dominio de frecuencias Bark mediante un banco de filtros conocido como Banco de Filtros Bark, que implementa la expresión:

$$\Omega[w] = 6 \cdot ln\left\{\frac{w}{1200\pi} + \left[\left(\frac{w}{1200\pi}\right)^2 + 1\right]^{0.5}\right\}, \tag{7}$$

donde ω es la frecuencia angular en rad/s.

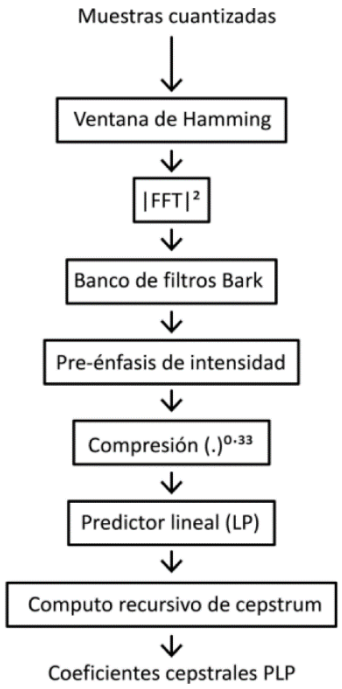

Figura 2

PROCEDIMIENTO PLP (TOMADA DE [2]).

El espectro de potencia mapeado P[ω] es convolucionado con el espectro de potencia de una curva que simula las bandas críticas de la audición humana descrita en [11]; esa convolución discreta procura muestrear las bandas críticas del espectro de potencia en intervalos de aproximadamente 1 bark. Luego de esta operación se logra que la primera muestra (asociada al 0 bark) y la última (asociada a la frecuencia de Nyquist) tengan igual valor que la muestra vecina más cercana. La última operación antes de la modelación de los coeficientes LP es la compresión raíz cúbica de amplitud. Con esta operación se intenta simular la relación no lineal entre la intensidad del sonido y la percepción de su tonalidad; logrando reducir la variación espectral de las bandas críticas del espectro permitiendo que la modelación LP pueda realizarse con modelos de bajo orden. La última operación del análisis PLP consiste en la aproximación mediante LPC de la compresión raíz cúbica. Los coeficientes autorregresivos obtenidos, como en la LPC convencional, pueden ser transformados a otro tipo de conjunto de parámetros como los coeficientes cepstrales. De esta manera pueden obtenerse un juego de coeficientes que codifiquen a

partir de la simulación psicofísica de la percepción humana información relevante para el RAH y lo suficientemente robustos para ser empleados como rasgos acústicos.

2.3.- TRANSFORMACIONES LINEALES

Normalización cepstral de media y varianza.
Las técnicas de normalización de rasgos tienen como objetivo reducir incongruencias entre los datos de entrenamiento y los datos de prueba [4] permitiendo: la creación de modelos más universales, robustecer los rasgos y elevar la efectividad del reconocimiento [4]. Estas técnicas de fácil implementación y bajo costo computacional han demostrado ser herramientas útiles y suelen incluirse en los sistemas de RAH, incluso en aquellos que enfrentan la tarea del reconocimiento en presencia de ruido [4]. En esta investigación fue utilizada la Normalización Cepstral de Media y Varianza (CMVN) que opera sobre dos momentos estadísticos para normalizar la secuencia de vectores de los datos: la media y la varianza. Luego de la normalización, la secuencia cepstral es de media cero y varianza unitaria.

Rasgos dinámicos de primer y segundo orden.
Los coeficientes cepstrales utilizados por los rasgos acústicos MFCC y los coeficientes PLP son rasgos estáticos [3], pues solo contienen información de una muestra dada. Al hablar, los órganos de fonación están cambiando su forma continuamente y este movimiento se refleja en el espectro en los cambios de las frecuencias y anchos de banda de los formantes, constituyendo, en algunos casos, un elemento identificativo del locutor y un elemento útil para el RAH. Bajo este principio es usual agregar información del contexto al vector de rasgo acústico de una muestra dada. Furui [12,13] propuso el uso de rasgos dinámicos que introdujeran información sobre el tránsito entre tramas, usando los coeficientes de primer orden de polinomios ortogonales obtenidos de cada coeficiente cepstral mediante:

$$\Delta c_t[n] = \left(\sum_{k=-K}^{K} \omega_k c_t + k[n] \right) \Big/ \left(\sum_{k=-K}^{K} \omega_k^2 \, (n = 1, \ldots, N) \right), \tag{8}$$

Estos coeficientes son una estimación temporal de la función derivada en el tiempo. El vector obtenido de (8) es conocido como cepstrum delta. El cálculo de segundas derivadas (delta-delta) pueden ayudar a mejorar el RAH [13,14].

Espectro Relativo.
La metodología Espectro Relativo (RASTA) es un tipo de transformación aplicada sobre los rasgos acústicos de la señal de habla observada desarrollada por H. Hermansky et al. [7] que pretende robustecer la PLP (y posiblemente a otras representaciones espectrales de corta duración) ante distorsiones espectrales lineales provocadas por el canal de comunicación [7]. Es entonces, RASTA, una técnica de normalización de canal implementada en forma de filtro cepstral. Esta metodología logra mejorar la precisión de las representaciones espectrales de las señales de habla cuando los datos de entrenamiento y los datos a reconocer son obtenidos desde diferentes micrófonos o canales [4]. RASTA combina el filtrado cepstral pasa-alto y pasa-bajo en una única función transferencial de respuesta al impulso infinita no causal (IIR) de la forma [7]:

$$H(z) = 0.1 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \cdot (1 - 0.98z^{-1})}, \tag{9}$$

La aplicación de la metodología RASTA sobre los coeficientes PLP se comporta de manera similar a la implementación en tiempo real de CMVN aplicado sobre MFCC, aunque con una tasa de error ligeramente mayor [14].

Análisis Discriminador Lineal.
El Análisis Discriminador Lineal (LDA) es un método ampliamente utilizado [4,15] en el reconocimiento de patrones para estimar un subespacio linear de los rasgos con buenas propiedades discriminatorias. En RAH esta tarea se traduce en encontrar una combinación lineal de rasgos acústicos que permita caracterizar los distintos fonemas producidos en la señal de habla observada; la ventaja de realizar esta tarea es que dicha combinación lineal mantendrá las propiedades discriminatorias del rasgo acústico reduciendo su dimensión.

La idea del LDA es encontrar una proyección de los datos donde la varianza entre las clases es grande en comparación con la varianza dentro de las clases. Bajo los supuestos de una distribución de clase Gaussiana y una matriz de covarianza intraclase común puede establecerse formalmente la idea del LDA como encontrar una matriz θ que maximiza el cociente:

$$J(\theta) = \frac{det(\theta \sum_b \theta^T)}{det(\theta \sum_\omega \theta^T)}, \quad (10)$$

donde ∑b es la matriz de covarianza inter-clases y ∑ω es la matriz de covarianza intra-clase. La solución a esta maximización es tomar los primeros vectores propios de la matriz $\sum^{-1}_w \sum b$ para una matriz de proyección dimensional p; que no es más que una matriz de mapeo que permite codificar cierta información contenida en una matriz de datos. LDA es invariante ante transformaciones lineales por lo que utilizar rasgos acústicos concatenados con transformaciones espectrales o temporales como entrada al LDA, solo implica elevar la complejidad y el costo computacional del análisis, producto del aumento en la dimensión de los rasgos acústicos, sin mejorar las características discriminatorias del resultado del LDA ni la efectividad del sistema de RAH.

Transformación Lineal de Máxima Verosimilitud.
La Transformación Lineal de Máxima Verosimilitud (MLLT) puede considerarse como un método de modelación de datos mediante matrices estructuradas de covarianza. MLLT es comúnmente usado en los sistemas dedicados al RAH para acercar las estimaciones estadísticas hechas sobre los datos a su verdadera distribución [6,16,17]. Esta aplicación de MLLT permite mejorar las estimaciones estadísticas hechas por los sistemas de RAH y de esta manera aumentar la efectividad de reconocimiento [16]. MLLT, es un modelo de estimación de covarianza que utiliza una representación estructurada para estimar las matrices de precisión. MLLT ofrece propiedades como la invariabilidad ante transformaciones lineales de los datos y una mejor descripción estadística de los vectores de rasgos acústicos. El único requerimiento que agrega MLLT es el de evaluar y sumar N × m exponenciales, donde N es el número de vectores de rasgos acústicos utilizados y m el número de parámetros usados. La MLLT sobre rasgos acústicos tienen un impacto considerable en la efectividad de reconocimiento de los sistemas de RAH, este hecho experimentalmente comprobado [16,17], la posiciona como una técnica a tener en cuenta en el diseño de sistemas de RAH.

3.- RECONOCIMIENTO AUTOMÁTICO DEL HABLA Y RUIDO.

Los sistemas utilizados en esta investigación pertenecen a una generación de sistemas dedicados al RAH que introduce el paradigma del reconocimiento de patrones y el uso de modelos estadísticos. Transformando el problema del RAH fundamentalmente en un problema de clasificación resuelto mediante un Clasificador de Bayes. Esta transformación implica que todos los errores de reconocimiento cometidos por los sistemas de RAH pertenecientes a esta generación son consecuencia de una clasificación errada [18]. Existen dos condiciones bajo las cuales la tasa de errores de clasificación incrementa en un Clasificador de Bayes:

- Condición Tipo 1, o Errores de Cotejo: ocurre debido a que las distribuciones de probabilidad usadas por el clasificador son estimaciones de las verdaderas distribuciones. Comúnmente resulta imposible caracterizar la distribución de probabilidad a posteriori de un conjunto de datos, por eso suele usarse un modelo que estime dicha densidad y entrenar sus parámetros con el conjunto de datos de entrenamiento para minimizar la divergencia entre los estimados y las verdaderas distribuciones [18]. A este tipo de errores de cotejo se les conoce como Errores de Cotejo del Modelo para diferenciarlos de los Errores de Cotejo de Datos. Los Errores de Cotejo de Datos son producidos cuando el valor de un rasgo utilizado para clasificar es modificado por alguna razón [18], comúnmente ruido. Resultando en una distribución de probabilidad a posteriori de los datos distinta a la asumida por la clasificación.
- Condición Tipo 2, o Incremento del Error de Bayes: Ocurre debido a la adición a la variable aleatoria a clasificar, x, dependiente de variable de clases c de una variable aleatoria, y, independiente a la clase (por ejemplo, ruido o variabilidades intrínsecas a la señal de habla como la coarticulación). Esta adición produce un incremento en el Error de Bayes cometido por el clasificador, o sea que si $z = x + y$ el Error de Bayes cometido por un clasificador basado en z será mayor que un clasificador basado solo en x.

En [18] se consideran tres factores claves, asociándolas a las condiciones Tipo 1 o Tipo 2, que afectan la efectividad de reconocimiento en los sistemas dedicados al RAH:

1) *La existencia de modelos estadísticos* que sustituyen las verdaderas probabilidades y distribuciones de probabilidad de los datos por estimaciones estadísticas de un conjunto de los datos. Este factor clave es una condición Tipo 1.

2) *La existencia de interferencias intrínsecas en las señales de habla*, tales como el énfasis usado por el locutor en ciertas sentencias o los fenómenos de coarticulación propias del habla, que no permiten crear generalizaciones en la clasificación de iguales sonidos emitidos por diferentes locutores. Estos atributos no lexicales pueden considerarse como variables aleatorias independientes, haciendo de la señal de habla observada una combinación de estas características con los rasgos lexicales. Este resultado es una condición Tipo 2, o sea, deriva en un incremento del Error de Bayes.
3) *La existencia de interferencias externas* tiene como resultado dos efectos: El primero es la introducción de Errores de Cotejo de Datos (Tipo 1), producidos por la diferencia entre las distribuciones de los datos empleados para clasificar y aquellos que deben ser clasificados. El segundo se debe a que, aunque la clasificación sea hecha con distribuciones apropiadas la existencia de estas interferencias producirá un incremento del Error de Bayes (Tipo 2). Este factor provoca el incremento de la tasa de clasificación tanto de Tipo 1 como de Tipo 2.

Los factores Dos y Tres son los de mayor interés para esta investigación ya que están relacionados directamente con el bloque Extractor de Rasgos, mientras el factor Uno está asociado únicamente a los modelos empleados en el bloque Decodificador. Los rasgos acústicos objeto de estudio de esta investigación, MFCC y los coeficientes PLP, están diseñados para lidiar con el factor Dos y una de los objetivos de esta investigación es determinar cuál de los rasgos acústicos estudiados permite a los sistemas dedicados al RAH lidiar mejor con el factor Tres.

4.- KALDI

El conjunto de herramientas utilizadas para la creación de los sistemas de RAH fue Kaldi [15]; el proceso de selección de este conjunto de herramientas utilizadas comenzó por la revisión bibliográfica referente a competencias internacionales de RAH como InterSpech o WOSSPA y a investigaciones de impacto en el área con la intención de identificar las herramientas para RAH más utilizadas [19-23]. De esta revisión solo se consideraron aquellas herramientas libres y de código abierto que ofrecían resultados relevantes, obteniendo tres candidatas: Hideen Markov Model Toolkit (HTK) (v3.4.1) [24], Sphinx-4 [25] y Kaldi [18].

Escapa a los objetivos planteados por esta investigación la comparación entre estos conjuntos de herramientas para el RAH. Por este motivo la investigación se ha servido del proyecto OASIS (*Open-source Automatic Speech recognition In Smart devices*) [25] y sus resultados comparando conjuntos de herramientas de código abierto para el RAH [26,27]. Los principales resultados descritos por el proyecto OASIS en la comparación entre HTK, Sphinx y Kaldi pueden resumirse en:

- Kaldi supera en efectividad de reconocimiento tanto a HTK como a Sphinx, debido a la implementación de las más novedosas técnicas de RAH como WFST, fMLLR, SGMM y DNN-HMM; pero con el costo computacional más elevado entre los tres modelos impactando en la eficiencia y los tiempos de reconocimiento.
- Sphinx ofrece la posibilidad de generar buenos resultados en poco tiempo. Su marco de trabajo incluye técnicas como LDA o MLLT, pero no ofrece otras más actuales como lo hace Kaldi.
- HTK es el conjunto de herramientas más complejo, en términos de configuración y uso; requiriendo mucho más conocimiento y esfuerzo que Sphinx y Kaldi. En cuanto a la efectividad del reconocimiento ofrece resultados similares a Sphinx.

Luego de esta revisión la herramienta seleccionada fue Kaldi. Los criterios seguidos para seleccionar a Kaldi como la herramienta a utilizar fueron la amplia y activa comunidad de desarrolladores e investigadores con que cuenta la herramienta, el hecho de que implementa las técnicas y métodos más novedosos del estado del arte en el área del RAH con los mejores índices de efectividad en reconocimiento y el interés del CENATAV en acumular experiencia sobre la creación de sistemas de RAH empleando Kaldi.

Conjunto de herramientas Kaldi

Kaldi es un conjunto de herramientas, libre y de código abierto, desarrollado por Daniel Povey y otros desarrolladores [15] para la investigación en el área del RAH. Kaldi permite la construcción de sistemas de RAH mediante una serie de rutinas de comandos shell bien documentadas. Los sistemas de RAH en Kaldi se basan en Transductores de Estados Finitos Ponderados (WFST) lo que permite optimizar los procesos de entrenamiento y decodificación. Kaldi está desarrollado en C++ y se lanzó bajo Apache License v2.0, que ofrece pocas restricciones, haciendo a esta herramienta ideal para el uso y desarrollo por comunidades de investigadores.

Kaldi depende de dos bibliotecas externas que pueden usarse libremente; la primera es OpenFst [28] utilizada para trabajar con el marco de los WFST y la segunda son las bibliotecas de álgebra numérica "Subrutinas Básicas de Álgebra Lineal"(BLAS) y el "Paquete de Álgebra Lineal"(LAPACK). En [15] D. Povey et. al. describen esquemáticamente mediante a Kaldi y agregan que el acceso a las funcionalidades de las bibliotecas del núcleo se proporciona a través de herramientas escritas en C++, que luego son llamadas desde un conjunto de comandos shell para construir y ejecutar el sistema de RAH.

Con el fin de evaluar el impacto de la selección del rasgo acústico, utilizado por un sistema de RAH, sobre la efectividad del reconocimiento, se procede a medir la WER obtenida por varios sistemas de RAH ante distintos escenarios de decodificación para los rasgos acústicos, MFCC y RASTA-PLP. Los sistemas utilizados en la experimentación serán entrenados en el mismo conjunto de entrenamiento, con el mismo Modelo de Lenguaje (ML) y sobre el mismo conjunto de datos de prueba; pero en cada sistema un nuevo Modelo Acústico (MA) y una transformación en el espacio de los rasgos será probada. Este procedimiento se realizará para ambos rasgos acústicos de interés a esta investigación, los MFCC y RASTA-PLP. Esto permitirá saber, si uno o ambos rasgos acústicos de interés potencian el rendimiento de los sistemas de RAH empleados en la experimentación. En la concepción de los sistemas se ha seguido un enfoque secuencial, aprovechando la posibilidad que ofrecen los sistemas de RAH basados en WFST de iniciar el entrenamiento de un nuevo MA con las mallas de palabras de un MA ya entrenado. La figura 3 muestra los MA empleados, la transformación en el espacio de los rasgos utilizada y la secuencia de construcción de los MA. Los sistemas referidos en la figura 13 comparten un mismo ML trifónico (de orden 3). El proceso de extracción de los rasgos acústicos seguido en esta investigación utilizando la herramienta Kaldi puede observarse en la figura 3.

Para la extracción de los MFCC se utilizó un Banco de Filtros Mel de 40 filtros (8 filtros por octava) [1], descartando el valor de la energía de la trama para un total de 13 coeficientes. En la extracción de RASTA-PLP se sustituyó el Banco de Filtros Bark propuesto por [2] por un Banco de Filtro Mel de 40 filtros (8 filtros por octava), con un orden del predictor lineal igual a 12 y un total de 13 coeficientes PLP. El nombre de cada sistema obedece a la nomenclatura utilizada por Kaldi donde cada sistema es nombrado por el tipo de MA que utiliza donde:

- mono: sistema HMM-GMM monofónico con 2500 estado y 15000 Mezclas Gaussianas (MG).
- tri1: sistema HMM-GMM trifónico con 2500 estados y 15000 MG.
- tri2: sistema HMM-GMM trifónico con 2500 estados y 15000 MG.
- tri3: sistema HMM-GMM trifónico con 2500 estados y 15000 MG, entrenado mediante Entrenamiento Adaptado al locutor.
- sgmm: sistema HMM-SGMM con 7000 estados y 9000 MG.
- sgmm+mmi: sistema HMM-SGMM con 7000 estados y 9000 MG, entrenado mediante Maximización de la Información Mutua.
- dnn: sistema híbrido HMM-DNN con 360 neuronas en la capa de entrada, con 2 capas ocultas de 5000 neuronas y con 2134 neuronas en la capa de salida.
- comb: arquitectura combinada a partir de los sistemas dnn y sgmm+mmi.

Debido a la incorporación en cada nuevo sistema de alguna característica con respecto al sistema anterior, ya sea en el modelo en sí o en la transformación hecha sobre los rasgos acústicos, es de esperar que el comportamiento de la WER sea de disminución mientras avanza la secuencia de sistemas. Este comportamiento será igual para ambos rasgos acústicos, pero al calcular la diferencia entre la WER de cada sistema en cada escenario y es apreciable que la diferencia entre WER se hace más negativa a medida que aumenta el ruido, de este hecho se llega a generalizaciones sobre qué rasgos ofrecen mejor robustez ante el ruido. En la figura 3 se relacionan todos los sistemas empleados, denotados por un rectángulo truncado.

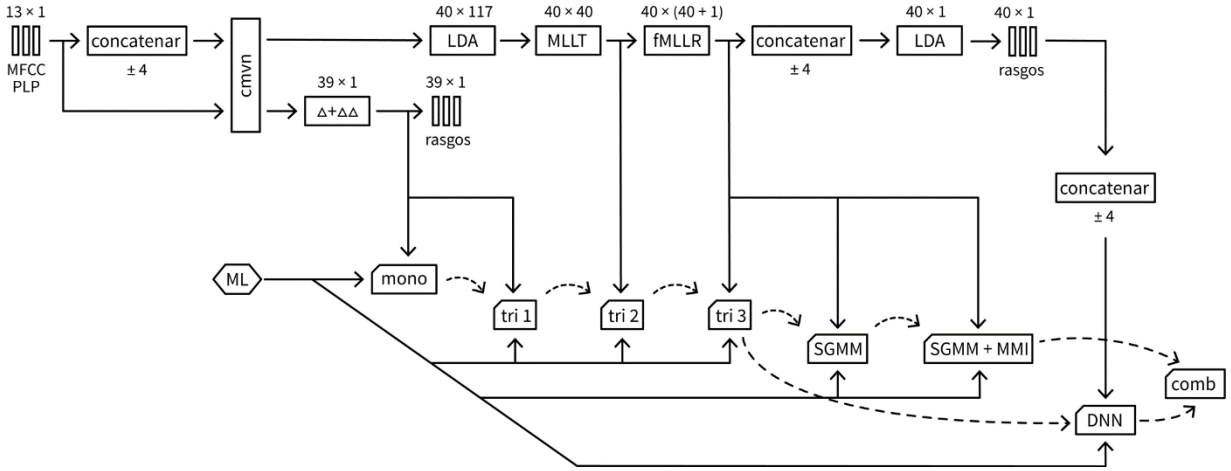

Figura 3

Secuencia de Sistemas de RAH empleados.

5.- EXPERIMENTOS

Base de datos TC-STAR.

El enfoque de RAH bajo el marco de los HMM implica poseer un conjunto de señales de habla y sus transcripciones. En el caso de esta investigación han sido empleados algunos conjuntos de la base de datos construida por el proyecto TC-STAR relacionadas en la **Tabla 1**. El proyecto TC-STAR, financiado por la Comisión Europea representa un esfuerzo a largo plazo centrado en la investigación avanzada en tecnologías del lenguaje como el RAH, el reconocimiento automático del locutor, la traducción automática del habla y síntesis del habla [29]. TC-STAR tiene como objetivo lograr avances tecnológicos que reduzcan significativamente la brecha entre el rendimiento de la traducción humana y la automática [30]. TC-STAR se centra en una selección de dominios de conversación sin restricciones, habla espontánea, en tres idiomas: inglés europeo, español europeo y chino mandarín [29]. Las grabaciones de TC-STAR utilizadas, que referenciadas a partir de ahora como TC-STAR-USADA, corresponden a sesiones del Parlamento Europeo o a sesiones de las Cortes Españolas donde los locutores solo hablen en español. Las sesiones grabadas de TC-STAR-USADA provienen de distintos escenarios (dos escenarios tipo auditorio), poseen múltiples locutores de ambos sexos (60 mujeres y 112 hombres) y diferentes edades, son sesiones de habla espontánea y debido a que algunos locutores intervienen en diferentes momentos del día o a veces en días distintos esto agrega variabilidad entre las sesiones grabadas de un mismo locutor haciendo sumamente complejo y real la tarea de RAH. Se siguió el criterio 90-10 para la creación de los conjuntos de entrenamiento y prueba; donde los datos de entrenamiento consisten en un conjunto de sesiones grabadas que correspondan al 90% del tiempo total (incluyendo silencios) de los datos de TC-STAR-USADA y los datos de prueba el 10% restante. La **Tabla 2** ofrece un resumen de las características de TC-STAR-USADA donde el término vocablo se refiere tanto a las palabras dichas como a las interjecciones y sonidos sin información lingüística etiquetados en la base de datos como ruido, tales como estornudos o carraspeos. El formato de los ficheros de audio es el estándar RIFF (.wav) codificados PCM, 16 bits con signo, a 16 kHz sin compresión.

Tabla 1
Relación de archivos TC-STAR utilizados

TCSTAR05	E0003	EVAL 05ES
TCSTAR06	E0012-01	CORTES06ES
	E0012-02	EPPS06ES-1
		EPPS06ES-1
TCSTAR07	E0026-01	CORTES07ES
	E0026	EPPS07ES

Tabla 2
Información sobre TC-STAR-USADA

	Datos de entrenamiento	Datos de prueba	Total
Sesiones	17163	1908	19071
Vocablos	241413	33492	274905
Vocablos Únicos	15722	4178*	16645
Locutores	153 | 53f 100m	23 | 9f 14m	172 | 60f 112m**
Horas	26h:38min:59s	3h:36min:16s	30h:16min:15s

(De los vocablos únicos de prueba sólo 918 no se encuentran entre los vocablos de entrenamiento y únicamente 2 locutores femeninos y 2 masculinos son compartidos entre los datos de entrenamiento y de prueba).

Normalización de textos.

Las bases de datos empleadas en la construcción de sistemas de RAH incluyen además de los ficheros de audio, ficheros de texto que contienen la transcripción manual de lo dicho en cada audio. Estos ficheros equivalen a las etiquetas de las clases en cualquier proceso de reconocimiento supervisado de patrones y son imprescindibles en RAH para la creación de modelos de lenguaje y de diccionarios fonéticos. En muchos casos los ficheros de transcripción solo contienen el nombre del fichero

de audio y su transcripción, elementos suficientes para herramientas como Kaldi; pero algunas bases de datos como TC-STAR ofrecen información adicional: la identificación del locutor, su sexo y los tiempos de inicio y fin de la locución. Esta información adicional puede ser utilizada por técnicas de entrenamiento adaptado, tales como SAT.

Contar con transcripciones que detallen información sobre los locutores e incluyan etiquetas de eventos acústicos sin información lingüística contribuye a elevar la efectividad de reconocimiento de los sistemas de RAH. Es imprescindible que esta información aparezca en las transcripciones en un estricto formato; pues puede ocurrir que, en el momento de transcripción manual, diferentes transcriptores asignaran etiquetas distintas a un mismo locutor ('Reguera_Díaz' y 'REGUERA_DIAZ') o que en algunos momentos decidieran transcribir la locución de fechas y números como palabras y en otros como números. Incluso aunque la transcripción fuera rigurosa y el proceso de estandarizar los ficheros de texto utilizados por los sistemas de RAH fuera adecuado es necesario un procesamiento sobre estos datos con el fin de eliminar sucesos que incrementen la variabilidad fonética en las transcripciones y por consecuencia del diccionario fonético y en el ML. Tales sucesos pueden ser el uso de abreviaturas, acrónimos y siglas en el proceso de transcripción manual que afectan las estimaciones estadísticas del ML.

El proceso sobre las transcripciones que resulta en la estandarización del formato de estas, recibe el nombre de Normalización de Textos, y en el caso de esta investigación cumple la función adicional de acondicionar las transcripciones a los requerimientos de la herramienta Kaldi, tales como: la necesidad de que todos los elementos de la transcripción tengan una única transcripción fonética, haciendo de los signos de puntuación un problema a resolver debido a que estos carecen de representación fonética y por tanto son imposibles de transcribir fonéticamente. Otro elemento inevitable es la existencia de ruidos de fondo que pueden alterar la calidad auditiva de las grabaciones, y por consiguiente la del sistema de reconocimiento en su conjunto. Asimismo, existen otros elementos producidos por los locutores, como las risas o los estornudos que, aunque no supongan información fonéticamente útil, sí necesitan ser modelados de alguna manera para asegurar que no entorpezcan el RAH. Es ideal modelar de forma independiente estos sonidos; pero en muchas ocasiones no aparecen debidamente señalados en las transcripciones o son señalados bajo una misma etiqueta.

A continuación, es descrito el proceso de Normalización de Textos empleado en esta investigación:

1) Extraer las transcripciones de los ficheros proporcionados por la base de datos TC-STAR eliminando la información restante (marcas de tiempo, número de canal, identificadores de locutores) de tal manera que cada línea del fichero resultante equivalga únicamente a la transcripción de una sesión de habla.
2) Asociar a las transcripciones de los sonidos sin información fonética etiquetas que puedan ser procesadas por la herramienta Kaldi. Por ejemplo, asociar la transcripción “<NOISED>” a la etiqueta fonética “NOISED”.
3) Eliminar los signos de puntuación existentes.
4) Eliminar espacios falsos (espacios dobles, triples o mayores).
5) Convertir todas las letras a mayúsculas conservando las tildes.
6) Convertir todos los números y fechas a palabras.
7) Sustituir todas las siglas y acrónimos por la transcripción de su pronunciación. Por ejemplo “PSOE” por “PESOE”.
8) Eliminación de símbolos o letras distintos a los del idioma español. Por ejemplo, tildes francesas o portuguesas en nombres propios de estos idiomas.

En general se recomienda la eliminación de tildes para evitar posibles errores ortográficos en las transcripciones; pero el procedimiento seguido en esta investigación aprovecha la riqueza prosódica del idioma español generando un Lexicón sensible a las fuerzas de pronunciación marcadas por las tildes. Esta decisión nos permite agregar información fonética a los modelos empleados por los sistemas de RAH creados; a expensas de asumir los posibles errores ortográficos de las transcripciones.

Base de ruidos y herramienta Fant.

En la sección 3 se estableció como fenómenos de interés a esta investigación en el RAH en presencia de ruido a los factores:

- Factor Dos: la existencia de interferencias intrínsecas en las señales de habla.
- Factor Tres: la existencia de interferencias externas.

El factor Dos está presente en la naturaleza de las locuciones, habla espontánea, de la base de datos TC-STAR-USADA. Sin embargo, el factor Tres debe ser simulado. Para esto fue necesario utilizar una base de ruidos y una herramienta que nos permitiera simular condiciones de ruido con distintos niveles de Relación Señal a Ruido (SNR) en las señales de los datos de prueba de TC-STAR-USADA para evaluar el impacto de la selección de MFCC o RASTA-PLP como rasgo acústico

utilizado por un sistema de RAH sobre la WER en escenarios ruidosos, o en otras palabras comparar la robustez de los rasgos acústicos ante distintos niveles de ruido

La base de datos de ruidos seleccionada en esta investigación ha sido DEMAND, en su versión 1.0 del 9 de junio del 2013 desarrollada por Joachim Thiemann, Nobutaka Ito, Emmanuel Vincent bajo la licencia Creative Commons Attribution-ShareAlike 3.0 Unported License [30]. DEMAND declara como objetivos proveer un conjunto de grabaciones que permita probar algoritmos en presencia de ruidos en ambientes reales obtenidos de diferentes configuraciones de un arreglo microfónico usando 16 canales, con distancias entre 5 cm y 21.8 cm entre micrófonos [30]. Estas características hacen de DEMAND una base de ruidos factible para la simulación de ruidos reales en señales de habla obtenidas de escenarios poco ruidosos mediante la adición de diferentes tipos de ruido y niveles de SNR; a diferencia de otras bases de ruidos (como AURORA, la base de datos de ruidos de fondo CHiME o la base de datos NOISEX) que consideran el uso de ambientes controlados. DEMAND ha sido dividida en 6 categorías de ruido, de las cuales cuatro son en interiores y dos al aire libre. Las sesiones en interiores han sido clasificadas en Domésticas, Oficinas, Público y Medios de Transporte; mientras que las grabadas al aire libre son Calle y Naturaleza. Cada categoría está dividida en tres ambientes de grabación. La **Tabla 3** muestra la clasificación de todos los grupos de ruidos y una descripción de estos.

Las grabaciones fueron hechas a 48 kHz durante un tiempo de 300 s. La duración de 300 s para cada grabación tiene como objetivo lidiar con la posibilidad (debido a que los ambientes no son controlados) de que ocurran ruidos que no formarán parte del ruido de fondo natural del ambiente y que en caso de ocurrir solo representarían una porción pequeña de la duración total de la grabación. Las grabaciones a usar están remuestreadas a 16 kHz. El arreglo microfónico empleado consiste en 16 micrófonos dispuestos en 4 filas espaciadas de tal manera que cada micrófono está a 5 cm de distancia de su vecino inmediato [31]. El arreglo durante las grabaciones quedó en el plano paralelo al suelo, montado en un trípode estándar que suspendía el arreglo a una distancia de 1.4 m del suelo [31]. El equipo usado consistió en 16 micrófonos capacitivos omnidireccionales Sony ECM-C110 [32] conectados al conversor A/D USB Inrevium/Tokyo Electron Device TD-BD-16ADUSB [32]. El conversor fue conectado a laptops con sistema operativo Windows o Linux. Las pistas fueron capturadas usando las herramientas provistas por el conversor USB y almacenadas en su formato por defecto (.ich), con una frecuencia de muestreo de 48 kHz y una ganancia de preamps de +20 dB. Para el cambio de formato por defecto (.ich) al estándar RIFF (.wav) se usó el script de MATLAB ich2wav.m provista por la base de datos [31].

La herramienta seleccionada para simular los ficheros de audio de los datos de prueba de la base de datos TC-STAR-USADA con los ruidos de la base de datos DEMAND fue FaNT – *Filtering and Noised Adding Tool* desarrollada por H. Guenter Hirsch [33]. Esta herramienta permite agregar ruido a sesiones de habla grabadas con una SNR deseada. De esta manera fueron creados los escenarios que simulan con bastante realismo distintos tipos de ruido según su nivel de SNR y de distinta naturaleza, debido a la utilización de todos los tipos de ruidos de la base de ruidos DEMAND.

6.- RESULTADOS

Definidas las bases de datos y las herramientas empleadas se procede en este acápite a la descripción de los escenarios de experimentación propuestos. Siguiendo la clasificación de ruidos en escenarios interiores y de ruidos en escenarios al aire libre dada en DEMAND, se procede a crear dos escenarios de experimentación llamados TC_STAR_IN_DOOR y TC_STAR_OUT_DOOR compuestos por todas las muestras de los datos de prueba de la base TC-STAR-USADA simuladas en el caso de TC_STAR_IN_DOOR con una selección aleatoria de todos los ruidos de escenarios interiores y en el caso TC_STAR_OUT_DOOR con una selección aleatoria de todos los ruidos de escenarios al aire libre. Como la herramienta Fant permite ajustar el nivel de la SNR con la que serán simuladas las señales de habla, fueron creados escenarios TC_STAR_IN_DOOR y TC_STAR_OUT_DOOR con valores de SNR que varían aleatoriamente en dos intervalos de valores (5-15 dB y de 15-20 dB). Esta metodología ha permitido crear seis escenarios con distintos niveles de SNR y tipos de ruido:

1. TC_STAR_DATA_AJUSTADA
2. TC_STAR_USADA
3. TC_STAR_IN_DOOR_5dB-15dB
4. TC_STAR_OUT_DOOR_5dB-15dB
5. TC_STAR_IN_DOOR_15dB-25dB
6. TC_STAR_OUT_DOOR_15dB-25dB

La nomenclatura empleada en cada escenario aclara el conjunto de ruidos de la base de ruidos DEMAND empleados para simular las señales de habla limpia de TC-STAR-USADA y el intervalo de valores de la SNR. Vale aclarar que el conjunto TC_STAR_DATA_AJUSTADA es una excepción en la nomenclatura empleada, debido que el ML utilizado por los sistemas de RAH de este escenario incluyen las transcripciones de los datos de prueba. El caso del escenario

TC_STAR_USADA, la nomenclatura hace alusión a que los datos de prueba de TC-STAR-USADA se han mantenido intactos.

Tabla 3
Descripción de la base de ruidos DEMAND.

Escenarios	Categorías	Ambientes	Descripción
Interiores	Domésticas	DWASHING	ruidos dentro de un baño
		DKITCHEN	ruidos dentro de una cocina durante la preparación de comidas
		DLIVING	ruidos dentro de la sala de una casa
	Público	PSTATION	ruidos en el área principal de transferencia en una estación de metro congestionada
		PCAFETER	ruidos en una cafetería de una oficina concurrida
		PRESTO	ruidos en un restaurante universitario en el horario de almuerzo
	Oficinas	OOFFICE	ruidos en una pequeña oficina con tres personas usando computadoras
		OHALLWAY	ruidos en un recibidor dentro de un edificio de oficinas, con personas o grupo de personas pasando ocasionalmente
		OMEETING	ruidos en una sala de reuniones durante una discusión
	Medios de transporte	TMETRO	ruidos en un metro
		TBUS	ruidos en un ómnibus público
		TCAR	ruidos en un automóvil privado
Aire Libre	Calle	STRAFFIC	ruidos en una intersección concurrida
		SPSQUARE	ruidos en una plaza pública con muchos turistas
		SCAFE	ruidos en una terraza de un café en una plaza pública
	Naturaleza	NFIELD	ruidos en un campo deportivo con actividades
		NRIVER	ruidos en un riachuelo
		NPARK	ruidos en un parque urbano concurrido

MFCC vs. PLP.

Los experimentos realizados en esta investigación fueron realizados a partir de siete sistemas de RAH cuyo bloque Extractor de Rasgos produce MFCC como rasgo acústico de la señal de habla observada, otros siete sistemas con el bloque Extractor de Rasgos que produce RASTA-PLP como rasgo acústico de la señal de habla observada y un sistema de RAH de arquitectura combinada. A partir de este punto las siglas PLP, en los nombres de los modelos, serán usadas para hacer referencia a RASTA-PLP. Todos los sistemas comparten un mismo conjunto de datos de entrenamiento: TC-STAR-USADA, consistente en señales de habla espontánea obtenidas en un escenario con condiciones acústicas favorables a la tarea de RAH (con bajo nivel de ruido); un mismo Corpus y por tanto un mismo ML (con excepción de TC-STAR-DATA-AJUSTADA). La diferencia radica en los MA, dentro de cada conjunto de sistemas. Los sistemas han sido entrenados siguiendo el algoritmo descrito en la figura **13**. Fueron entrenados, entonces, dos conjuntos de ocho sistemas que utilizan como rasgo acústico transformaciones sobre los MFCC o RASTA-PLP como muestra la **Tabla 4**.

La experimentación consistió en medir la WER de los sistemas creados en la decodificación (obtención de la transcripción) de los datos de prueba en los seis escenarios propuestos. Como línea base de la experimentación se declaran las WER obtenidas en la decodificación en el escenario TC_STAR_USADA mostradas en **Tabla 5**. Los tipos de errores por omisión (DEL), sustitución (SUB) o inserción (INS) de palabras son referidos en porcientos y no en cantidad de errores cometidos para observar la distribución de los tipos de errores por escenario.

Un comportamiento de posible generalización a ambos grupos de sistemas es la disminución de la WER a medida que se complejizan los MA, con la excepción en la secuencia: tri2_plp, tri3_plp, sgmm_plp. Este comportamiento no resulta de mucho interés pues se debe más a las mejoras introducidas por los nuevos MA que a un impacto en la WER debido a la selección de un tipo de rasgo acústico. Por otro lado, debido a la pequeña diferencia en la WER (menor de un 1% con excepción de tri2_mfcc con tri2_plp que es del 5.09%) de sistemas con igual MA, pero basados en rasgos acústicos distintos es posible afirmar que en escenarios de bajo nivel de ruido la selección de MFCC o RASTA-PLP como rasgo acústico a emplear por el sistema no ofrece un impacto considerable en la WER obtenida en la decodificación.

Los resultados de la **Tabla 5** ofrecen además de la WER de las transcripciones de cada sistema, la distribución de los errores cometidos siendo las sustituciones (SUB) por marcada diferencia el error más frecuente en todos los modelos. Este hecho pudiera ser atenuado por un ML que contará con más realizaciones de secuencias de palabras ocurridas en las señales a decodificar, de tal manera que se contara con hipótesis de estas secuencias de palabras y se pudiera decidir mejor reduciendo la cantidad de sustituciones, contribuyendo así a tener hipótesis de decodificación más robustas.

Tabla 4
Transformaciones de los rasgos por sistema.

Sstemas	Modelo Acústico	Rasgo
mono_mfcc	HMM-GMM monofónico	$MFCC + CMVN + \Delta + \Delta\Delta$
tri1_mfcc	HMM-GMM trifónico	$MFCC + CMVN + \Delta + \Delta\Delta$
tri2_mfcc	HMM-GMM trifónico	$MFCC \pm 4 + CMVN + LDA + MLLT$
tri3_mfcc	HMM-GMM trifónico	$MFCC \pm 4 + CMVN + LDA + MLLR + fMLLR$
sgmm_mfcc	HMM-SGMM trifónico	$MFCC \pm 4 + CMVN + LDA + MLLR + fMLLR$
sgmm+mmi_mfcc	HMM-SGMM trifónico	$MFCC \pm 4 + CMVN + LDA + MLLR + fMLLR$
dnn_mfcc	HMM-DNN trifónico	$MFCC \pm 4 + CMVN + LDA + MLLR + fMLLR \pm 4 + LDA$
mono_plp	HMM-GMM monofónico	$PLP + CMVN + \Delta + \Delta\Delta$
tri1_plp	HMM-GMM trifónico	$PLP + CMVN + \Delta + \Delta\Delta$
tri2_plp	HMM-GMM trifónico	$PLP \pm 4 + CMVN + LDA + MLLT$
tri3_plp	HMM-GMM trifónico	$PLP \pm 4 + CMVN + LDA + MLLR + fMLLR$
sgmm_plp	HMM-SGMM trifónico	$PLP \pm 4 + CMVN + LDA + MLLR + fMLLR$
sgmm+mmi_plp	HMM-SGMM trifónico	$PLP \pm 4 + CMVN + LDA + MLLR + fMLLR$
dnn_plp	HMM-DNN trifónico	$PLP \pm 4 + CMVN + LDA + MLLR + fMLLR \pm 4 + LDA$

Una evidencia a favor de esta hipótesis son los resultados de la **Tabla 6**, donde se relacionan las WER obtenidas del escenario TC_STAR_AJUSTADA. En el escenario de la **Tabla 6** el Corpus de entrenamiento contiene las transcripciones de los datos de prueba. De estos resultados vale hacer notar la drástica reducción de la WER, por debajo de un 10% para todos los modelos, como una ventaja de la inclusión de transcripciones de secuencias frecuentes en el ML. Sin embargo, no deberían esperarse grandes disminuciones de la WER sobre todos los datos de prueba ya que en la práctica raramente se tienen transcripciones exactas de todo lo que se espera decodificar. Sin embargo, es posible lograr una disminución considerable en la WER sobre aquellas locuciones que son probables ocurran y que por tanto pueden ser incluidas sus probables transcripciones en el Corpus de entrenamiento. Un ejemplo sería agregar textos periodísticos a un Corpus de entrenamiento de un sistema de RAH empleado para subtitular programas noticiosos, ya que los textos agregados contienen secuencias de palabras de muy probable emisión en los datos a transcribir.

Tabla 5
WER para escenario TC_STAR_USADA.

TC_STAR_USADA					
Sistemas	**WER**	**ERRORES**	**INS**	**DEL**	**SUB**
mono_mfcc	44,12%	14781	7%	31%	62%
tri1_mfcc	**30,20**%	10117	11%	25%	65%
tri2_mfcc	27,95%	9364	12%	25%	64%
tri3_mfcc	25,48%	8537	14%	23%	64%
sgmm_mfcc	**22,76**%	7626	15%	22%	63%
sgmm+mmi_mfcc	**21,28**%	7130	14%	23%	63%
dnn_mfcc	23,51%	7878	13%	25%	63%
comb_mfcc	**20,77**%	6957	13%	25%	62%
mono_plp	**44,09**%	14773	7%	30%	63%
tri1_plp	30,39%	10182	11%	25%	64%
tri2_plp	**22,86**%	7659	14%	23%	63%
tri3_plp	**25,35**%	8493	12%	25%	63%
sgmm_plp	22,86%	7659	14%	23%	63%
sgmm+mmi_plp	21,40%	7168	15%	22%	63%
dnn_plp	**22,97**%	7695	13%	23%	63%
comb_plp	21,05%	7053	12%	25%	62%

Otro resultado de la **Tabla 6** a tener en cuenta es la diferencia en la distribución de los porcientos de los tipos de errores, en especial los porcientos de omisiones (DEL) y SUB, con un incremento del orden del 20% en DEL (con un máximo de 29% para mono_plp y dnn_plp) y un decrecimiento del 20% en SUB (con un máximo de 26% para mono_plp y un mínimo de 18 para tri3_mfcc). Mientras que para los porcientos de errores por inserciones (INS) solo experimentaron variaciones de entre un 4% y un 1%, despreciables si son comparados con los porcientos de errores de DEL y SUB. Esta redistribución en los tipos de errores puede ser entendida para el caso de la disminución de errores por sustitución de palabras como una mejoría en la decodificación entre hipótesis similares (el modelo se "confunde" menos en el reconocimiento). Mientras que, en el caso del incremento de errores por omisión de palabras, debido a los valores tan pequeños en la WER, puede inferirse que son producidos por aquellas palabras de difícil reconocimiento debido a las condiciones acústicas relacionadas con su locución (ruido ambiental, brevedad de la palabra, mala articulación) y por tanto son imposibles de corregir mediante mejores hipótesis de decodificación.

Las tablas: **Tabla 7** y **Tabla 8**, contienen los resultados de las decodificaciones de los sistemas de RAH utilizados en escenarios ruidosos. La nomenclatura empleada en los escenarios ofrece información sobre la relación señal-ruido a la que fueron ensuciadas los datos de entrenamiento de TC-STAR-USADA. Vale señalar que los ruidos pertenecientes a la categoría aire libre, codificada en la nomenclatura de los escenarios como OUT_DOOR, tiene un impacto más agresivo en la degradación de la calidad perceptiva de las señales de habla que los ruidos pertenecientes a la categoría interiores, codificada en la nomenclatura de los escenarios como IN_DOOR, con idénticos niveles de SNR. Es de esperar, entonces, mayores WER en los escenarios con ruido de aire libre y es de esperar que sean los más reveladores en cuanto a la robustez de los rasgos acústicos.

Es notable el hecho de que para todos los sistemas en todos los escenarios las WER menores son obtenidas por aquellos sistemas que utilizan como rasgo acústico de la señal de habla a los MFCC. Una medida de la robustez interesante entre los rasgos resulta de calcular las diferencias de las WER entre los sistemas que utilizan como rasgo acústico de la señal de habla los MFCC y los que utilizan los coeficientes RAST-PLP. La **Tabla 9** muestra estos resultados para cada escenario, donde los valores negativos significan WER inferiores para los sistemas con rasgo acústico MFCC y los valores positivos WER inferiores para los sistemas con rasgo acústico RASTA-PLP.

Tabla 6
WER para escenario TC_STAR_USADA.

TC_STAR_DATA_AJUSTADA					
Sistemas	**WER**	**ERRORES**	**INS**	**DEL**	**SUB**
mono_mfcc	9,21%	3086	4%	59%	37%
tri1_mfcc	4,44%	1489	9%	49%	42%
tri2_mfcc	4,53%	1519	12%	43%	45%
tri3_mfcc	4,46%	1494	13%	41%	46%
sgmm_mfcc	3,27%	1096	12%	47%	41%
sgmm+mmi_mfcc	3,12%	1044	12%	46%	42%
dnn_mfcc	2,97%	996	8%	54%	38%
comb_mfcc	86,32%	28909	0%	2%	98%
mono_plp	9,06%	3037	4%	59%	37%
tri1_plp	4,82%	1616	9%	49%	42%
tri2_plp	4,59%	1538	11%	45%	44%
tri3_plp	4,41%	1479	13%	42%	45%
sgmm_plp	3,27%	1096	12%	47%	41%
sgmm+mmi_plp	3,12%	1044	12%	46%	42%
dnn_plp	3,05%	1022	9%	53%	38%
comb_plp	86,33%	28923	0%	2%	98%

La **Tabla 9** ordena los escenarios atendiendo a la intensidad de ruido, colocando en la primera columna el escenario con mayor relación señal-ruido y con la categoría de ruido menos agresivo y en la última el escenario con menor relación señal-ruido y con la categoría de ruido más agresiva. De la **Tabla 9** puede concluirse que el comportamiento general de las diferencias entre las WER es aumentar proporcionalmente con la intensidad del ruido existente en el escenario; este comportamiento puede ser utilizado como una evidencia a favor de la selección de los MFCC como rasgo acústico a emplear en escenarios ruidosos. Vale hacer notar como hecho interesante que los sistemas tri2 ofrecen las mayores diferencias en cada escenario y entre los escenarios con menor y mayor nivel de SNR.

Tiempos y Espacio de cómputo.

Las tablas: **Tabla 5**, **Tabla 7**, **Tabla 8** y **Tabla 9**, ofrecen información sobre las WER obtenidas por los sistemas en los distintos escenarios de decodificación. Está información resulta muy valiosa para los objetivos propuestos en esta investigación, pero teniendo en cuenta que en muchos casos el tiempo consumido por los sistemas dedicados al RAH se desea menor o igual al tiempo de la señal de habla que se procesa; por tal motivo la **Tabla 10** muestra los tiempos consumidos por cada proceso durante el RAH. Estos tiempos fueron realizados en una laptop con sistema operativo Linux Mint 18 “Sarah”, con microprocesador Intel Core i3-380M @ 2.53 GHz x 4 y dos 2 GB de memoria RAM.

Tabla 7
WER para escenarios TC_STAR_IN_DOOR.

	TC_STAR_IN_DOOR_5dB-15dB					TC_STAR_IN_DOOR_15dB-25dB				
Sistemas	WER	ERRORES	INS	DEL	SUB	WER	ERRORES	INS	DEL	SUB
mono_mfcc	57,29%	19194	5%	31%	64%	46,49%	15577	7%	29%	64%
tri1_mfcc	41,71%	13974	7%	30%	63%	31,50%	10553	10%	26%	64%
tri2_mfcc	40,05%	13418	8%	31%	62%	29,40%	9850	10%	26%	64%
tri3_mfcc	34,50%	11557	10%	28%	62%	26,36%	8833	12%	24%	64%
sgmm_mfcc	32,21%	10790	11%	26%	63%	24,32%	8149	14%	23%	63%
sgmm+mmi_mfcc	30,51%	10221	11%	27%	63%	22,87%	7662	13%	23%	63%
dnn_mfcc	30,79%	10312	10%	27%	63%	24,09%	8072	13%	24%	63%
comb_mfcc	28,82%	9652	10%	29%	61%	22,24%	7451	12%	27%	62%
mono_plp	60,20%	20170	4%	35%	61%	47,91%	16052	7%	29%	64%
tri1_plp	45,49%	15241	7%	32%	61%	32,79%	10984	11%	24%	65%
tri2_plp	45,99%	15407	6%	36%	58%	30,91%	10356	10%	26%	64%
tri3_plp	37,77%	12654	9%	31%	60%	27,23%	9122	12%	24%	64%
sgmm_plp	34,95%	11709	11%	27%	62%	24,94%	8356	12%	26%	63%
sgmm+mmi_plp	33,70%	11289	11%	28%	61%	23,57%	7897	14%	22%	63%
dnn_plp	34,31%	11494	8%	32%	59%	24,84%	8323	11%	26%	63%
comb_plp	31,89%	10685	9%	31%	59%	22,85%	7654	13%	24%	63%

Tabla 8
WER para escenarios TC_STAR_OUT_DOOR.

	TC_STAR_OUT_DOOR_5dB-15dB					**TC_STAR_OUT_DOOR_15dB-25dB**				
Sistemas	**WER**	**ERRORES**	**INS**	**DEL**	**SUB**	**WER**	**ERRORES**	**INS**	**DEL**	**SUB**
mono_mfcc	**61,75**%	20689	4%	34%	62%	**47,17**%	15805	6%	30%	64%
tri1_mfcc	**46,84**%	15693	6%	32%	62%	**31,95**%	10703	9%	27%	64%
tri2_mfcc	**45,26**%	15162	6%	36%	59%	**29,70**%	9949	11%	24%	65%
tri3_mfcc	**36,47**%	12220	8%	31%	61%	**26,37**%	8836	12%	24%	64%
sgmm_mfcc	**33,77**%	11315	10%	28%	62%	**24,16**%	8095	13%	24%	63%
sgmm+mmi_mfcc	**32,47**%	10879	8%	29%	61%	**22,93**%	7683	13%	24%	63%
dnn_mfcc	**34,29**%	11484	9%	30%	61%	**24,08**%	8066	12%	25%	63%
comb_mfcc	**30,73**%	10293	9%	31%	60%	**21,71**%	7275	12%	26%	62%
mono_plp	64,44%	21589	4%	34%	62%	48,51%	16252	5%	32%	63%
tri1_plp	50,75%	17004	5%	34%	61%	33,49%	11219	10%	25%	65%
tri2_plp	52,86%	17709	5%	40%	56%	31,78%	10648	9%	27%	64%
tri3_plp	40,84%	13682	7%	33%	60%	27,51%	9217	11%	25%	63%
sgmm_plp	37,48%	12557	10%	29%	61%	24,97%	8367	12%	25%	63%
sgmm+mmi_plp	36,42%	12203	8%	33%	59%	23,68%	7932	14%	23%	64%
dnn_plp	36,82%	12335	8%	34%	59%	25,06%	8396	12%	24%	63%
comb_plp	34,35%	11508	7%	36%	57%	22,86%	7659	11%	27%	62%

Tabla 9

Diferencia de WER entre MFCC y PLP.

	TC_STAR USADA	**TC_STAR IN_DOOR 15dB-25dB**	**TC_STAR OUT_DOOR 15dB-25dB**	**TC_STAR IN_DOOR 5dB-15dB**	**TC_STAR OUT_DOOR 5dB-15dB**
Sistemas					
mono	0,03%	-1,42%	-1,34%	-2,91%	-2,69%
tri1	-0,19%	-1,29%	-1,54%	-3,78%	-3,91%
tri2	5,09%	-1,51%	-2,08%	-5,94%	-7,60%
tri3	0,13%	-0,87%	-1,14%	-3,27%	-4,37%
sgmm	-0,10%	-0,62%	-0,81%	-2,74%	-3,71%
sgmm+mmi	-0,12%	-0,70%	-0,75%	-3,19%	-3,95%
dnn	0,54%	-0,75%	-0,98%	-3,52%	-2,53%
comb	-0,28%	-0,61%	-1,15%	-3,07%	-3,62%

La primera consideración al leer los datos de la **Tabla 10** tiene que ver con el hecho de que los tiempos de decodificación son los más importantes a tener en cuenta, pues el entrenamiento de los modelos es totalmente independiente al proceso de decodificación. Lo usual es entrenar y almacenar el modelo entrenado (entrenamiento *offline*) y luego decodificar un proceso que ocurre en tiempo real (decodificación *online* o en línea). Entrenar *offline* y decodificar *online* es lo deseable, pero para que sea posible es necesario que los tiempos de decodificación sean menores o similares al tiempo de las señales de habla que se desean codificar. Vale recordar que el tiempo total de los datos de prueba utilizados es de 3h:36min:16s por lo que, como muestra la **Tabla 10**, solo es posible decodificar *online* con los modelos tri2 y tri3, pues los tiempos de decodificación de estos modelos son inferiores al del conjunto de señales a decodificar. Aunque los tiempos de decodificación de los sistemas tri2 y tri3 permitan emplearlos como sistemas de decodificación *online* debe tenerse en cuenta que las WER obtenidas por estos sistemas son mayores que las de los sistemas que le preceden. Resultando en un compromiso entre la efectividad del reconocimiento y el tiempo necesario para ejecutarlo. Otro elemento a destacar de la **Tabla 10** es que todos los procesos relacionados con los sistemas de RAH, con excepción del ML, que utilizan RASTA-PLP consumen menos tiempo que sus análogos con MFCC como rasgo acústico.

La **Tabla 11** muestra las diferencias entre los tiempos empleados por los procesos utilizando como rasgo acústico los MFCC y RASTA-PLP. Aunque las mayores diferencias se observan entre los tiempos de entrenamiento de los modelos más complejos, sgmm+mmi y dnn; las diferencias más importantes a tener en cuenta deben ser entre los sistemas tri2 y tri3, debido a que son los sistemas que permite la decodificación *online*. En el caso de tri2 la diferencia temporal es de dieciséis minutos, que comparada con el tiempo de los datos a decodificar puede ser ignorada; al contrario de los cincuenta minutos de diferencia para los sistemas tri3 donde el empleo de RASTA-PLP marca una diferencia importante en los tiempos de decodificación.

Aunque las mayores diferencias entre los tiempos de decodificación están asociadas a los sistemas con MA tri1 y mono, no son relevantes pues estos sistemas suelen emplearse como sistemas de inicialización y no para decodificar *online* debido a las altas WER que ofrecen y a los tiempos necesarios para crearlos.

Los requerimientos de memoria impuestos por los rasgos tienen las siguientes fuentes: ancho, en milisegundos, de la ventana utilizada; números de filtros por octava usados y la cantidad de coeficientes cepstrales. En nuestra experimentación ambos rasgos fueron utilizados: ventanas de 10ms; 40 filtros Mel o Bark, dependiendo del rasgo y 12 coeficientes cepstrales. Estas características seleccionadas garantizan condiciones iguales para ambas representaciones e implican, como consecuencia, iguales requerimientos de memoria. En cuanto a las transformaciones lineales delta-delta, LDA o MLLT, que buscan en el primer agregar información contextual valiosa a la representación o en el segundo caso reducir la dimensionalidad de la representación o como en el tercero mejorar la representación estadística del conjunto de entrenamiento a costo de elevar el costo computacional y el volumen de memoria necesario, fueron usadas para ambos rasgos idénticas configuraciones. Ambos rasgos acústicos y las transformaciones lineales aplicadas sobre ellos, debido a la necesidad de evaluar su desempeño en función del enfoque que utilizan, fueron usados con configuraciones idénticas y de esta forma los requisitos de memoria que imponen son también idénticos.

Tabla 10

Tiempo empleado por sistema y tarea

	Extractor de Rasgos		0:16:12	
	Modelo de Lenguaje		0:00:18	
	Entrenamiento		Decodificación	
	MFCC	PLP	MFCC	PLP
mono	4:20:59	**4:15:06**	16:20:59	**15:03:37**
tri1	1:16:09	**1:09:56**	8:35:14	**6:42:10**
tri2	1:30:16	**1:26:03**	2:07:47	**1:51:16**
tri3	1:59:56	**1:46:56**	3:16:03	**2:25:17**
sgmm	14:29:30	**13:52:47**	6:28:22	**5:11:09**
sgmm+mmi	2:50:37	**22:28:03**	6:50:44	**5:30:10**
dnn	0:08:23	**21:10:18**	9:24:04	**7:58:28**
comb	-	-	9:45:21	**8:19:50**

Tabla 11

Diferencias temporales entre sistemas

	Diferencias Temporales	
Extracción de Rasgos	00:00:04	
	Entrenamiento	Decodificación
mono	00:05:53	01:17:22
tri1	00:06:13	01:53:04
tri2	00:04:13	00:16:31
tri3	00:13:00	**00:50:46**
sgmm	00:00:00	00:26:27
sgmm+mmi	**04:22:34**	00:03:21
dnn	**02:58:05**	00:05:02
comb	-	00:00:05

7.- CONCLUSIONES

La investigación presentada en torno al problema de selección entre los MFCC y los coeficientes PLP como rasgo acústico empleado en la tarea del RAH en escenarios ruidosos arribó a la siguiente generalización empírica: los MFCC son rasgos acústicos más robustos que los coeficientes PLP ante la tarea del RAH en presencia de ruido. Las conclusiones principales de la investigación son:

- Los MFCC son más robustos ante la tarea del RAH en presencia de ruido que los coeficientes PLP.
- MFCC y los coeficientes PLP ofrecen WER similares ante la tarea del RAH en escenarios sin ruido o poco ruidosos.
- El RAH es más rápido si se utilizan los coeficientes PLP.
- Ambos rasgos acústicos, MFCC y los coeficientes PLP, imponen iguales restricciones de memoria durante el proceso de RAH.
- El conjunto de herramientas Kaldi permitió implementar un entorno de experimentación extenso y complejo utilizando las técnicas y métodos más novedosos del estado del arte en el área del RAH.

- Incluir en el Corpus de entrenamiento secuencias de palabras específicas hará más efectivo su reconocimiento automático.

El valor de esta investigación consiste en que aporta evidencias sobre la robustez de los MFCC como rasgo acústico empleado en el RAH en escenarios ruidosos.

REFERENCIAS

1. Davis SB, Mermelstein P.; Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans on ASSP; 1980; 4(5): 357-366
2. Hermansky H. Perceptual linear predictive analysis of speech. J Acoust Soc Am. 1990; 87(4):1738-1752.
3. Peinado AM, Segura JC. Speech Recognition with HMMs in *Speech Recognition Over Digital Channels: Robustness and Standard*s. John Wiley & Sons Ltd; 2006.
4. Droppo J, Acero A. Environmental Robustness. In: J. Benesty MMS, Yiteng Huang, editor. Springer Handbook of Speech Processing. Berlin: Springer; 2008. p. 653-677.
5. Pylkkönen J. LDA Based Feature Estimation Methods for LVCSR. International Conference on Spoken Language Processing Interspeech 2006. p. 389-392, Pittsburgh, Pennsylvania, USA.
6. Gales MJF, editor. Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition, Computer, Speech & Language, 1998, 12:75-98
7. Hermansky H., Morgan N., Bayya A., Kohn P.. RASTA-PLP speech analysis technique. Proc IEEE Int Conf Acoust ICASSP-92, 1992, San Francisco, CA, USA.
8. Povey D, Yao K. A basis representation of constrained MLLR transforms for robust adaptation. Comput Speech Lang. 2012; 26:35-51.
9. Miyajima C., Watanabe H., Kitamura T., Katagiri S., Speaker Recognition Based on Discriminative Feature Extraction – Optimization of Mel-Cepstral Features Using Second-Order All-Pass Warping Function. Proc Eurospeech 6th; 1999.
10. Schafer RW. Homomorphic Systems and Cepstrum Analysis of Speech. In: Benesty J, Sondhi MM, Huang Y, editors. Springer Handbook of Speech Processing. Berlin: Springer; 2008. pp. 161-80.
11. Fletcher H. Audittory patterns. Rev Mod Phys. 1940; 12:47-65.
12. Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Trans on ASSP. 1981:254-272.
13. Furui S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans on ASSP. 1986; 34:52-59.
14. Hanson B. A., Applebaum T.H. Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with lombard and noisy speech. Proc of ICASSP; 1990. Alburquerque, NM, USA.
15. Mariani J, et. al. Survey of the State of the Art in Human Language Technology. Cole R, editor. Cambridge: Cambridge University Press and Giardini; 1997.
16. Olsen PA, Ramesh AG, editors. Extended MLLT for Gaussian Mixture Models. Transactions in Speech and Audio Processing; 2001.
17. Psutka JV. Benefit of Maximum Likelihood Linear Transform (MLLT) Used a Different Levels of Covariance Matrices Clustering in ASR Systems. In: Matoúsek V, Mautner P, editors. TSD 2007; Berlin: Springer-Verlag; 2007. pp. 431-438.
18. Young SJ. HMMs and Related Speech Recognition Technologies. In: J. Benesty MMS, Yiteng Huang, editor. Springer Handbook of Speech Processing. Berlin: Springer; 2008. pp. 539-555.
19. Povey D, et. al, editors. The Kaldi Speech Recognition Toolkit. IEEE Workshop on Automatic Speech Recognition and Understanding; 2011.
20. Kim C, Stern RM, editors. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Transactions on audio, speech, and language processing; 2016.
21. Tachioka Y, Watanabe S, Hershey JR, editors. Effectiveness of discriminative training and feature transformation for reverberated and noisy speech. Proc ICASSP; 2013.
22. Alam M.J., O'Shaughnessy D., Kenny P. A novel feature extractor employing regularized MVDR spectrum estimator and subband spectrum enhancement technique. 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), 2013. Algiers, Algeria.
23. Umit HY, Satya D, editors. Perceptual MVDR-based cepstral coefficients (PMCCS) for robust speech recognition. Proc ICASSP; 2003.

24. Tachioka Y, et. al. Prior-based binary masking and discriminative methods for reverberant and noisy speech recognition using distant stereo microphones. Journal of Information processing. 2017; 25:407-16.
25. Young SJ, et. al. The HTK Book [web page]. Cambridge: Cambridge University Engineering Department; 2006.
26. Lamere P, et. al., editors. The CMU SPHINX-4 Speech Recognition System. Proc ICASSP; 2003.
27. Stuttgart B-WCSU. OASIS—Open-Source Automatic Speech Recognition In Smart Devices. [web page] Germany: Baden Wuerttemberg Cooperative State University Stuttgart; 2014 [cited 2018]; Available from: http://www.dhbw-stuttgart.de/themen/kooperative-forschung/fakultaet-technik/oasis.html.
28. Gaida C., et. al. Comparing Open-Source Speech Recognition Toolkits. Proc. ICSLP; 2014.
29. Allauzen C, et. al, editors. OpenFst: a general and efficient weighted finite-state transducer library. Proc CIAA; 2007.
30. ELRA-ELDA. TC-STAR, ELDA. [web page] Spain: ELRA-ELDA; 2000 [cited 2018]; Available from: http://www.elra.info/en/projects/archived-projects/tc-star.
31. Thiemann J, Ito N, Vincent E, editors. The Diverse Environments Multi-Channel Acoustic Noise Databas (DEMAND): A database of multichannel environmental noise recordings. Proc of Meetings on Acoustics ICA2013; 2013.
32. Sony. (2000). Sony [web page]. Available: https://www.sony.es/electronics/support/audio-video-accessories microphones/ecm-cs3/specifications [Accessed: 25 Feb. 2018].
33. Hark. (2000). Hark [web page]. Available:https://www.sony.es/electronics/support/audio-video-accessories-microphones/ecmcs3/specifications [Accessed: 25 Feb. 2018].
34. Hirsch HG. FaNT: filtering and noise adding tool. Niederrhein University of Applied Sciences. [web page] Germany: Niederrhein University of Applied Sciences; 2005 [cited 2018 May.]; Available from: http://dnt.kr.hsnr.de/download.

AUTORES

José Manuel Ramírez Sánchez, Ingeniero en Telecomunicaciones y Electrónica, CENATAV-DATYS, La Habana, Cuba. Intereses de investigación: Reconocimiento automático del habla y detección de términos hablados. E-mail: jsanchez@cenatav.co.cu.

Ana Montalvo Bereau, Licenciada en Física, Investigadora Agregada del CENATAV-DATYS, La Habana, Cuba. Intereses de investigación: Identificación del idioma hablado, procesamiento de habla, detección de términos hablados, aprendizaje profundo, representación del habla. E-mail: amontalvo@cenatav.co.cu.

José Ramón Calvo de Lara, Ingeniero en Telecomunicaciones. Doctor en Ciencias técnicas, Investigador Titular, CENATAV-DATYS, La Habana, Cuba. Intereses de investigación: Identificación del idioma hablado, procesamiento de habla, detección de términos hablados, aprendizaje profundo, representación del habla. E-mail: jcalvo@cenatav.co.cu.