Propuesta metodológica para la construcción de un sistema ASR basado en HMM usando diseño de experimentos

Fidelia Milena Rubio Ramos
Julian Yezid Olarte Ramos
Laboratorio de Automática, Microelectrónica e Inteligencia Computacional - LAMIC
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia
ingfideliarubio@gmail.com, julian@naturasoftware.com

Marco Aurelio Alzate
Jairo Soriano
Laboratorio de Automática, Microelectrónica e Inteligencia Computacional - LAMIC
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia
malzate@udistrital.edu.co, jsoriano@udistrital.edu.co

***Resumen*— En este artículo se presenta una propuesta metodológica para construir un sistema de reconocimiento automático de voz basado en Modelos Ocultos de Markov. A partir de la variación de algunos parámetros del sistema de ASR, se implementaron diferentes modelos en Sphinx-4 y se utilizaron técnicas de diseño de experimentos para determinar cuál de ellos presentaba mejores resultados. Los indicadores de desempeño utilizados fueron la tasa de error de reconocimiento de palabras (WER por sus siglas en inglés), el tiempo de procesamiento y el uso de memoria durante el reconocimiento.**

El sistema de reconocimiento se encuentra orientado a hablantes andino-orientales de nacionalidad Colombiana y se encuentra limitado a un conjunto de 20 palabras relacionadas con un sistema de marcado telefónico automático.

Reconocimiento Automático de Voz, Sphinx, Modelos Ocultos de Markov, Diseño de experimentos.

I. INTRODUCCIÓN

Un sistema de reconocimiento automático del habla (ASR por sus siglas en inglés) busca a través de un modelo de cómputo, extraer palabras desde una señal acústica y representarlas en forma de texto [1,2]. Con este fin, el algoritmo de reconocimiento intenta representar la señal basado en sus características acústicas y sus relaciones con símbolos fonéticos [3]. Cuando estos sistemas son utilizados en presencia de ruido, con palabras que no se encuentran dentro de su vocabulario o ante su uso por parte de múltiples hablantes, se presentan bajos niveles de precisión en el reconocimiento [4-6].

Por otra parte, la teoría del diseño de experimentos permite realizar pruebas a un sistema de forma rigurosa. Los métodos del diseño experimental utilizados en aplicaciones de diseño de ingeniería permiten evaluar y comparar diferentes configuraciones de un sistema para determinar cuál presenta un mejor desempeño en función de una variable de respuesta [7].

El propósito de este trabajo es proponer un método para construir un sistema de reconocimiento automático de voz basado en Modelos Ocultos de Markov (HMM por sus siglas en inglés) para hablantes Colombianos. El método propuesto plantea la implementación de diferentes modelos a partir de la variación de algunos parámetros del sistema y el uso de técnicas de diseño de experimentos y análisis de datos para determinar cuál de ellos presenta un mejor desempeño según tres indicadores: WER, tiempo de procesamiento y uso de memoria.

En la sección dos se presentan las técnicas de ASR utilizadas en la propuesta: coeficientes cepstrales (MFCC) para la etapa de extracción de características, modelos de n-gramas para el modelo de lenguaje y HMM para la etapa de modelado acústico y reconocimiento. En la tercera sección, se presenta una breve descripción de la teoría básica del diseño de experimentos y análisis de datos requerida para la propuesta metodológica planteada.

En la sección cuatro se describe la herramienta de implementación de los modelos y se presenta la propuesta metodológica para la construcción del sistema de reconocimiento automático del habla. Posteriormente se describe la implementación y la evaluación de los diferentes modelos de ASR a partir de la variación de ciertos parámetros de configuración del sistema y el análisis de los resultados obtenidos. Finalmente, en la sección cinco se presentan las conclusiones generadas a partir del desarrollo del trabajo y del análisis de los resultados obtenidos.

II. RECONOCIMIENTO AUTOMÁTICO DE VOZ

Generalmente, el proceso de reconocimiento de voz puede ser dividido en tres partes: extracción de características, clasificación y selección [3,8,9]. La señal de voz es capturada y pre-procesada. Luego, pasa a través de un bloque que obtiene las características relevantes para el reconocimiento y las introduce a un sistema de clasificación. Éste asigna un valor a cada una de las posibles salidas y finalmente el bloque de selección, determina la que considera más adecuada.

En la etapa de pre-procesamiento, deben considerarse dos elementos: la presencia/ausencia de voz y el uso de funciones de ventana [3,1,8].

En la etapa de extracción de características, se pretende parametrizar una señal de audio en una secuencia de características relevantes para el reconocimiento. Para su realización, se han propuesto diferentes técnicas que comprenden el análisis estadístico de la señal y el análisis en el dominio de la frecuencia, entre otros. [3,10,11].

Una de estas técnicas, denominada análisis Cepstral se basa en un modelo de generación de voz en el que la señal se compone de una secuencia de excitación e(n) convolucionada con la respuesta al impulso del modelo del sistema vocal θ(n) [3,10]. El cepstrum representa una transformación de la señal de voz con las siguientes propiedades [1]:

• Los componentes de la señal de voz se separan en el cepstrum.
• Los componentes de la señal de voz son combinados linealmente en el cepstrum.

El cepstrum de una señal de voz s(n) se encuentra definido por la siguiente ecuación:

$$c_s(n) = \mathcal{F}^{-1}\{log|\mathcal{F}\{s(n)\}|\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|S(w)|e^{jwn}dw \quad (1)$$

Donde *F{.}* es la transformada de Fourier en tiempo discreto. Si los coeficientes cepstrales son obtenidos a partir de la escala Mel, estos son denominados coeficientes Mel-Cepstrum.

Finalmente, en la etapa de clasificación y selección, se asigna un valor a la similitud de las características obtenidas en el bloque anterior con características de referencia y se determina cuál de ellas es la más similar [12]. Para esto se utilizan técnicas como los HMM.

A. Modelos Ocultos de Markov

En estos modelos la observación es una función probabilística del estado. El modelo resultante es un proceso estocástico doblemente embebido con un proceso estocástico subyacente que puede ser observado a través de otro conjunto de procesos estocásticos que producen la secuencia de observaciones [1,3,10,8].

En la Figura 1 se representa un HMM de izquierda a derecha de cinco estados. El valor observado en cada uno de los estados obedece a una función de distribución de probabilidad presente en cada estado [4].

Un HMM se caracteriza por lo siguiente:

• El número de estados en el modelo (N). Los estados independientes se denotan *{1, 2, ..., N}* y un estado en un tiempo t se denota q_t.
• El número de símbolos diferentes por estado (M).
• La distribución de probabilidad de transición entre estados (a_{ij}).
• La distribución de probabilidad de observación de un símbolo ($B = \{b_j(k)\}$).
• La distribución de estados inicial ($\pi = \pi_i$).

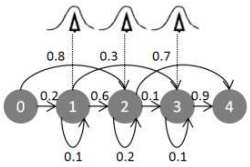

Figura 1. HMM de izquierda a derecha de cinco estados.

Una especificación compacta de un HMM se puede realizar de la siguiente forma:

$$\lambda = (A, B, \pi) \quad (2)$$

Con valores apropiados de N, M, A, B y π, un HMM puede ser usado como un generador de una secuencia de observaciones:

$$O = (o_1 o_2 \dots o_T) \quad (3)$$

En la cual cada observación o_T es uno de los símbolos contenidos en *V* y *T* es el número de observaciones en la secuencia.

B. Tres problemas básicos para los HMM

Dado un HMM hay tres problemas básicos de interés a ser solucionados:

1. Evaluación de probabilidad. Dada una secuencia de observaciones $O = (o_1 o_2 \dots o_T)$ y un modelo $\lambda = (A, B, \pi)$ cómo calcular de forma eficiente $P(O|\lambda)$, la probabilidad de la secuencia de observación, dado el modelo.
   Para solucionar este problema, se emplean los algoritmos de Forward o Backward [3,10].
2. Aplicación. Dada una secuencia de observaciones $O = (o_1 o_2 \dots o_T)$ y un modelo $\lambda=(A,B,\pi)$ cómo elegir una secuencia de estados $q = (q_1 q_2 \dots q_T)$ que es óptima en algún sentido.
   El algoritmo de Viterbi [3,10] es una técnica que permite encontrar la secuencia de estados que maximiza $P(q|O,\lambda)$. Este se aplica a cada observación de la palabra y así se obtiene una partición de las observaciones dado que se conoce desde qué estado se ha producido cada una de ellas.
3. Entrenamiento. Cómo ajustar los parámetros del modelo $\lambda = (A, B, \pi)$ para maximizar $P(O|\lambda)$. Se debe ajustar los parámetros del modelo para satisfacer cierto criterio de optimización. Para ello se selecciona $\lambda = (A, B, \pi)$ tales que su probabilidad $P(O|\lambda)$, sea localmente maximizada utilizando el método Baum-Welch [3,10].

C. Aplicación de los HMM a los sistemas de ASR

Dada una cadena de símbolos que representan una palabra, se busca la correspondiente cadena de símbolos del diccionario de pronunciación [13].

Sea O una entrada acústica que corresponde a una secuencia de símbolos de un alfabeto:

$$O = o_1, o_2, \dots, o_t \quad (4)$$

Y la cadena de palabras *W*:

$$W = w_1, w_2, \dots, w_t \quad (5)$$

Se tiene la siguiente función probabilística:

$$W = arg \max_{W \in L} P(W|O) \quad (6)$$

Usando el teorema de Bayes y considerando que *P(O)* es igual para todas las secuencias de palabras posibles, la cadena de palabras más probable dada una secuencia de observaciones se calcula por medio de la siguiente expresión [4]:

$$W = arg \max_{W \in L} P(O|W)P(W) \quad (7)$$

Donde *O* es la señal de voz observada, *P(O|W)* y *P(W)* son la probabilidad asociada a la secuencia *W* generada por el modelo acústico y el modelo de lenguaje respectivamente.

Esto es, la frase más probable *W* dada una secuencia de observación *O* puede ser calculada por medio del producto de la probabilidad a priori de que la palabra *W* haya sido pronunciada *P(W)* dada por el modelo de lenguaje y la probabilidad de una secuencia de observaciones acústicas condicionadas por una palabra *P(O|W)* generada por el modelo acústico. En la figura 2 se muestra la forma en que se relacionan estos elementos.

D. Modelos de lenguaje

Un modelo estadístico de lenguaje provee una estructura de lenguaje a nivel de palabras por medio de grafos o modelos estocásticos de n-gramas [11]. El objetivo de estos modelos es proveer una estimación de la probabilidad de una secuencia de palabras W para el reconocimiento a partir de un corpus de texto, para lo que se trata el proceso de generación de palabras como un proceso de Markov [3] [11].

El conjunto de datos seleccionado para la construcción del modelo de lenguaje puede ser modificado haciendo uso de herramientas como el uso de claves de contexto y la aplicación de técnicas de suavizado de los valores de probabilidad calculados.

Las claves de contexto son marcadores representados por símbolos que indican eventos tales como los límites de frases y párrafos [14].

Las técnicas de suavizado consisten en reducir la probabilidad estimada para eventos observados y redistribuirla entre los eventos no observados (n-gramas que no ocurren en el texto de entrenamiento) [11,14]. Supóngase que *N* sea el tamaño de un texto de muestra y que n_r es el número de palabras (m-gramas) que ocurre en el texto *r* veces

$$N = \sum_r r n_r \quad (8)$$

La estimación de máxima verosimilitud de la probabilidad de ocurrencia de un evento está dada por

$$P_{ML} = \frac{r}{N} \quad (9)$$

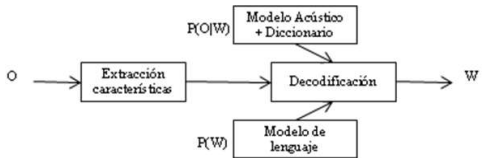

Figura 2. **Elementos básicos de un sistema reconocedor**

En una muestra dispersa, la estimación de máxima verosimilitud tiende a asignar valores altos de probabilidad a los datos observados y el valor cero a los no observados. Para corregir dicha tendencia, se toma parte de la masa de probabilidad de los eventos observados menos de *k* veces por medio de un coeficiente de descuento *dr* y se redistribuye en los eventos no observados [14].

$$P_T = \frac{r^*}{N} \quad (10)$$

Donde

$$r^* = r d_r \quad (11)$$

Según la definición del coeficiente de descuento, se definen diferentes métodos de descuento:

- Descuento de Good-Turing.

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}} \quad (12)$$

Para $1 \leq r \leq k \; y \; d_r = 1 \; para \; r > k$

- Descuento Lineal.

$$d_r = 1 - \alpha \quad (13)$$

- Descuento de Witten-Bell.

$$d_r = \frac{R}{R+t} \quad (14)$$

Donde R es el número de eventos que se presentan en el conjunto de datos de entrenamiento.

Luego de construir un modelo de lenguaje a partir de un corpus de entrenamiento, se debe realizar su validación con un conjunto de datos. Para esta etapa de evaluación se utiliza el concepto de perplejidad del modelo de lenguaje. La perplejidad (*B*) es una medida asociada a la entropía que representa la incertidumbre promedio que el sistema tiene cuando va a reconocer una palabra según el modelo de lenguaje [11]. Cuanto menor sea su valor, mejor será el modelo.

$$\hat{B} = 2^{\hat{H}} = \hat{P}(w_1, w_2, \dots, w_Q)^{-1/Q} \quad (15)$$

III. DISEÑO DE EXPERIMIENTOS

El diseño de experimentos y análisis de datos cuenta con metodologías estadísticas bien establecidas para seleccionar la estrategia experimental más adecuada en cada caso y evaluar los resultados garantizando un valor específico o rango de confiabilidad en los resultados obtenidos [7,15].

En [7,15] Montgomery propone una guía para la planeación y diseño de los experimentos que permite utilizar un enfoque

estadístico en el análisis de los resultados. El procedimiento propuesto consta de los siguientes pasos:

- Planeación previa del diseño experimental. Definición del título del experimento, objetivos, factores de control, factores que se deben mantener constantes y variable de respuesta.

- Elección del diseño experimental. Selección del diseño experimental según el tamaño de la muestra, el número de factores y niveles y el uso de restricciones sobre la aleatorización.

- Realización del experimento. Verificación de que el experimento se realice tal como fue planeado.

- Análisis estadístico de los datos. Los datos obtenidos con la ejecución de los experimentos deben ser analizados por medio de métodos estadísticos que permitan obtener conclusiones objetivas acerca del nivel de confianza de un enunciado.

- Conclusiones y recomendaciones. Luego del análisis estadístico de los datos se deben generar conclusiones prácticas sobre los resultados obtenidos.

A. Análisis estadístico de los datos

Según el tipo de experimento realizado y la naturaleza de los datos, se pueden utilizar diferentes métodos de análisis estadístico de los resultados. A continuación se describen brevemente los métodos utilizados en el presente trabajo.

Análisis de Varianza (ANOVA) [7]

Este análisis evalúa la hipótesis de que los k grupos provienen de la misma población o de poblaciones con la misma media.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a \quad (16)$$

$$H_1: \textit{El enunciado anterior no es verdadero para al menos una } \mu_i$$

El procedimiento de análisis de varianza consiste en obtener dos estimaciones para la varianza (entre y dentro de los grupos) y comparar estas estimaciones utilizando estadísticos que siguen una distribución F. Si la varianza entre los grupos es grande respecto a la varianza dentro de los grupos, probablemente se debe rechazar la hipótesis nula.

Comparaciones múltiples de Tukey [7]

Comparaciones entre los tratamientos con el fin de determinar diferencias específicas. Cuando los tamaños de muestras son iguales se utiliza la prueba de Tukey.

$$H_0: \mu_i = \mu_j \quad (17)$$

$$H_1: \mu_i \neq \mu_j, \quad \forall\, i \neq j$$

ANOVA para k muestras relacionadas

Este tipo de análisis se utiliza cuando todos los tratamientos se aplican a todas las unidades experimentales para controlar la variabilidad entre los grupos. El modelo utilizado es el siguiente:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad (18)$$

Donde τ_i es el efecto del tratamiento i-ésimo y β_j es un parámetro asociado con el sujeto j-ésimo. La hipótesis nula especifica que no hay ningún efecto de los tratamientos:

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0 \quad (19)$$

$$H_1: \textit{Al menos una } \tau_i \textit{ es diferente de cero}$$

Comparaciones por pares [7]

Comparación por pares de medias utilizando la prueba t de dos muestras. Esta prueba requiere que se valide el supuesto de normalidad y se utiliza cuando se tienen varianzas diferentes y desconocidas.

$$H_0: \mu_i = \mu_j \quad (20)$$
$$H_1: \mu_i \neq \mu_j$$

Si se desea realizar k comparaciones por pares de medias de muestras diferentes se debe utilizar la corrección de Bonferroni.

Análisis de varianza de Friedman [16]

Es una prueba estadística no paramétrica que permite evaluar la hipótesis nula de que las muestras pertenecen a la misma población. Los datos pueden ser obtenidos a partir de un diseño completamente aleatorio o a partir de un diseño con medidas repetidas. Este análisis evalúa la hipótesis de que los k grupos provienen de poblaciones con la misma mediana θ.

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_k \quad (21)$$

$$H_1: \textit{El enunciado anterior no es verdadero para al menos una } \theta_i$$

IV. Propuesta del Sistema de Reconocimiento Automático de Voz

A partir de las técnicas estudiadas para los sistemas de reconocimiento automático de voz y de la teoría del diseño de experimentos como herramienta de comparación entre diferentes modelos del sistema de ASR, se genera una propuesta metodológica para la construcción de un sistema de reconocimiento automático de voz.

En la propuesta se plantean dos experimentos principales. Con el primero se pretende seleccionar el mejor modelo de lenguaje a partir de la evaluación de la perplejidad de diferentes modelos implementados. Con el segundo, se busca determinar cuál de los modelos implementados del sistema de ASR presenta un mejor desempeño en función del WER, tiempo de procesamiento y uso de memoria.

En la figura 3 se presenta la propuesta metodológica mencionada. En color café y azul se describen los pasos relacionados con el primer y segundo experimento respectivamente.

Inicialmente (Figura 3a) se realiza un planteamiento inicial del experimento del sistema de ASR, donde se define el título del experimento, sus objetivos y los posibles factores. Posteriormente, se definen los criterios de diseño de la base de datos y se realiza su construcción.

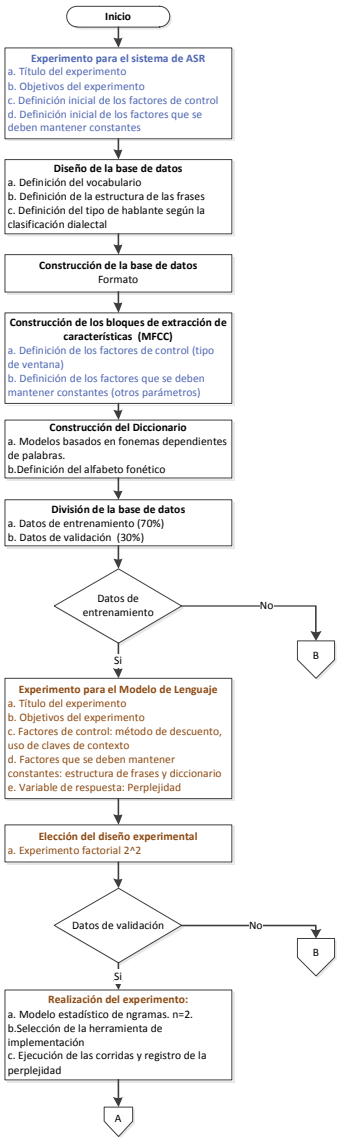

Figura 3a. Propuesta para la construcción del sistema de ASR.

El siguiente paso consiste en el diseño y configuración de los diferentes bloques que conforman el sistema de reconocimiento: se diseña y configura el bloque de extracción de características y luego se construyen los elementos de los bloques de clasificación y selección:

• Diccionario: basado en un alfabeto fonético, contiene las palabras definidas durante el diseño de la base de datos (Figura 3a).

• Modelo de lenguaje: considerando que se puede realizar una medición del desempeño de forma independiente del sistema de ASR a través de la perplejidad y con el objetivo de disminuir el número de tratamientos para el experimento del sistema de reconocimiento, se plantea un experimento independiente que permita definir cuál es el mejor modelo de lenguaje (Figura 3a).

Para ello, se define un nuevo experimento con título, objetivos, variable de respuesta (perplejidad) y factores de control a partir de los cuales se implementan diferentes modelos de lenguaje haciendo uso de los datos de entrenamiento. Posteriormente, con los datos de validación se ejecutan las corridas y se registran los valores de perplejidad obtenidos para cada uno de los modelos.

Se realiza el análisis estadístico de los datos como se muestra en el diagrama de flujo y finalmente se determina cual de los modelos es el que presenta un mejor desempeño definido como menor valor de perplejidad (Figura 3b).

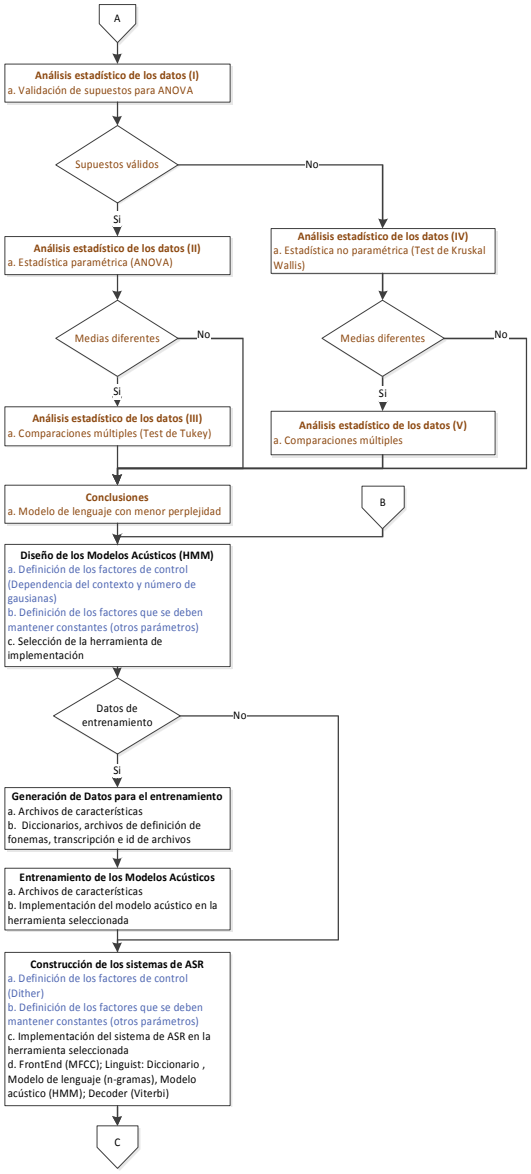

Figura 3b. Propuesta para la construcción del sistema de ASR.

• Modelo acústico: en este paso se revisan y definen nuevamente los factores y niveles relacionados con el modelo acústico para el experimento de ASR. Se realiza la generación y entrenamiento de los modelos acústicos. Se deben configurar e implementar diferentes modelos acústicos, uno para cada una de las combinaciones de los niveles de los factores definidos para el bloque de extracción de características y los niveles de los factores definidos para el modelo acústico (Figura 3b).

Luego, se revisa todo el planteamiento del experimento del sistema de reconocimiento y se procede con la ejecución de las corridas con los datos de validación. Durante cada corrida se registra el WER, el tiempo de procesamiento y el uso de memoria. Finalmente, se realiza el análisis estadístico de los datos en función de cada una de las variables de salida de tal forma que el experimento se divide en tres experimentos con variable de salida diferente y en cada caso, se encuentra el mejor sistema en función de la característica medida: WER, tiempo de procesamiento y uso de memoria (Figura 3c).

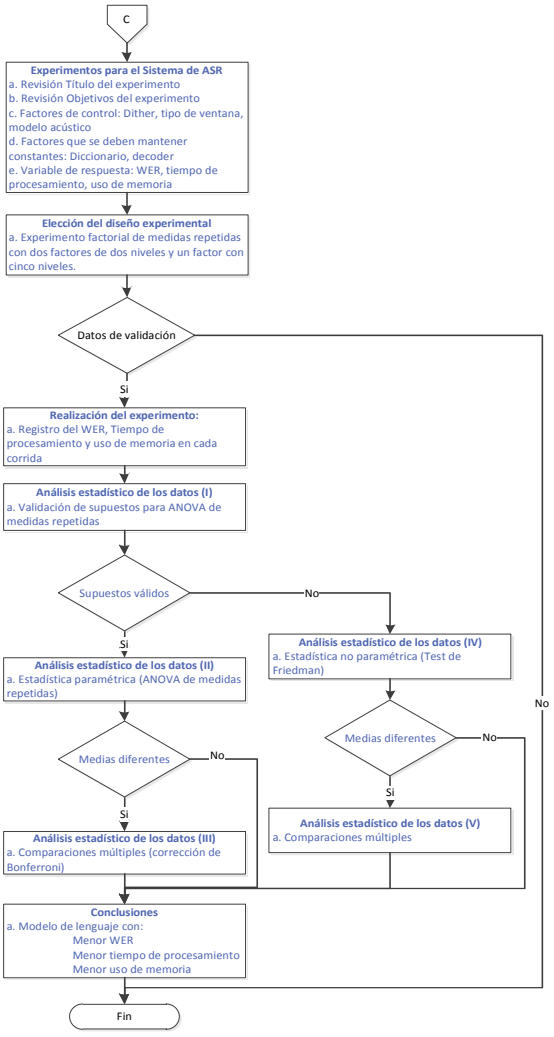

Figura 3c. Método propuesto para la construcción del sistema de ASR.

A. Herramienta de implementación

La implementación de la propuesta de ASR se realiza utilizando la herramienta Sphinx-4, una herramienta desarrollada en la Universidad de Carnegie Mellon basada en la construcción de Modelos Ocultos de Markov [17].

Los módulos que componen la herramienta de implementación son los siguientes [18]:

- FrontEnd. Toma las señales de audio de entrada y las parametriza en una secuencia de características.

- Linguist. Construye el grafo de búsqueda usando la estructura del lenguaje representada por un modelo y la estructura topológica de un modelo acústico. El linguist también utiliza un diccionario para mapear las palabras del modelo de lenguaje en secuencias del modelo acústico. Tiene los siguientes componentes:

- *Modelo de lenguaje*. Provee la estructura del lenguaje al nivel de palabras.

- *Diccionario*. Provee pronunciaciones para las palabras encontradas en el modelo de lenguaje.

- *Modelo acústico*. Provee un mapeo entre una unidad de voz y un HMM que puede ser clasificado entre un grupo de características provenientes del FrontEnd.

Típicamente, el linguist divide cada palabra del vocabulario en una secuencia de unidades de palabras dependientes del contexto. Pasa las unidades y su contexto al modelo acústico, recuperando los HMM asociados con esas unidades. Cada HMM produce una calificación para una característica observada.

- Grafo de búsqueda. Representa el espacio de búsqueda. Es una estructura de grafo producida por el linguist de acuerdo a algún criterio, utilizando conocimiento del diccionario, el modelo acústico y el modelo de lenguaje.

- Decoder. Toma las características generadas por el FrontEnd y el grafo de búsqueda generado por el linguist para realizar la decodificación y obtener los resultados.

B. Captura y preprocesamiento de la señal

La creación del corpus para la elaboración del modelo acústico, del vocabulario y del modelo de lenguaje se enmarca en un sistema de marcado telefónico automático. La selección de la aplicación obedece a que ésta contiene un vocabulario clásico para sistemas de reconocimiento de habla continua, como son los diez dígitos de cero a nueve.

Para elaborar el corpus se utilizó un vocabulario de 15 palabras diferentes que corresponden a la información típica requerida cuando se realiza una llamada y cinco nombres, cuatro de hombre y uno de mujer. La organización de las oraciones se realizó siguiendo una estructura típica de marcado telefónico. Los números telefónicos utilizados fueron tomados del directorio telefónico del año 2010 de la ciudad de Bogotá.

1) Recolección de datos

Debido a que el dialecto genera variaciones de la señal de voz, la propuesta de reconocimiento de voz está dirigida

hablantes de un dialecto específico del castellano de Colombia. La propuesta de sistema se orienta a múltiples hablantes del idioma español de la región Andino-Oriental según la clasificación dialectal propuesta en [19].

Además del subdialecto existen otros factores como el género y la edad del hablante que también generan variaciones en la señal de voz, por lo que la base de datos se construyó con hablantes hombres, mujeres y niños (menores de 12 años).

Cada uno de los hablantes leyó las diez frases que le fueron asignadas y luego las grabó de tal forma que en las grabaciones se escucharan como habla continua. La grabación fue realizada con las siguientes características (soportadas por Sphinx-4):

- Formato: wav
- Canales: 1 (mono)
- Tasa de muestreo: 16 KHz
- No. De bits por muestra: 16 bits
- Codificación: PCM

2) Extracción de Características

El método de extracción de características propuesto para el sistema reconocedor consiste en la extracción de los coeficientes Mel-Cepstrum de la señal (MFCC). Esta selección obedece a que según la literatura, en presencia de múltiples hablantes ésta técnica presenta mejores resultados en reconocimiento que las otras [20].

El primer bloque de la figura 4 corresponde a separación de la señal en segmentos de 10ms para su posterior procesamiento, luego cada segmento es procesado por una función de ventana. Los cuatro bloques siguientes corresponden a los pasos necesarios para la generación de los coeficientes Mel-Cepstrum. Las funciones de ventana y los parámetros de los filtros son variados dentro del experimento que se describe en la sección cinco. Finalmente, se extraen la primera y segunda derivada de los coeficientes para obtener el vector de características que ingresará al sistema de clasificación:

$$c_{m,n} = \{c_1, c_2, \dots, c_{13}, dc_1, dc_2 \dots, dc_{13}, d^2c_1, d^2c_2, \dots, d^2c_{13}\} \quad (15)$$

3) Clasificación

El sistema de reconocimiento propuesto se basa en modelos ocultos de Markov debido a que es la técnica más utilizada en el campo del reconocimiento de voz [21-24].

Modelo de Lenguaje. Fue creado utilizando el toolkit de modelamiento estadístico de lenguaje de la Universidad Carnegie Mellon. Los datos recolectados fueron divididos en dos conjuntos: 70% para el entrenamiento y 30% para la validación del sistema por medio del cálculo de la perplejidad [25].

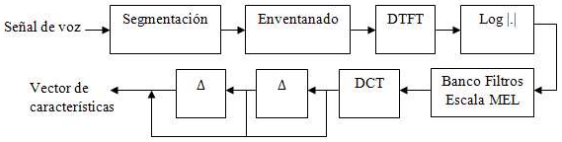

Figura 4. Extracción de características

Modelo Acústico. Se basa en modelos ocultos de Markov de tres estados y de tipo continuo, debido a que los modelos discretos aún cuando tienen un costo computacional moderado, discretizan el espacio de observaciones produciendo pérdidas de información que pueden deteriorar el rendimiento del sistema [26]. Como la propuesta se orienta a un sistema de habla continua y debido al reducido tamaño de la base de datos de entrenamiento, se propone el uso de modelos basados en palabras y modelos basados en fonemas dependientes de palabras.

Los datos recolectados fueron divididos en dos conjuntos: 70% para entrenamiento y 30% para validación.

Diccionario. Debido a que se propone el uso de modelos basados en fonemas dependientes de palabras, el diccionario se encuentra compuesto de las 20 palabras del vocabulario y su transcripción fonética.

El alfabeto fonético utilizado es *Worldbet*. Este alfabeto codifica los símbolos del alfabeto fonético internacional (IPA por sus siglas en inglés) y otros símbolos adicionales para cada lenguaje [27].

4) Decodificación

La decodificación se realiza utilizando el algoritmo de Viterbi, que determina la secuencia de estados correspondiente a una cadena de símbolos observada.

C. Experimentación y Análisis de Resultados

Se realizaron dos experimentos, con el primero se pretende seleccionar el modelo de lenguaje que se va a utilizar y con el segundo se busca realizar la validación de los sistemas de reconocimiento automático de voz planteados y determinar la propuesta que presente mejores resultados.

El enfoque estadístico y el procedimiento seguido para el diseño de los experimentos es el descrito en la primera parte de la sección 3.

a) Modelo de lenguaje

a. Planeación previa al experimento:

Título del experimento: Estudio de técnicas para la generación de un modelo de lenguaje para un sistema de reconocimiento de voz para un vocabulario limitado a 20 palabras.

Objetivos del experimento: Cuantificar la perplejidad de diferentes propuestas de modelo de lenguaje generadas a partir de la información encontrada en la revisión de la literatura. Determinar la influencia de las técnicas y sus interacciones en la perplejidad y determinar cuál modelo de lenguaje presenta un mejor desempeño.

Factor	Nivel	Descripción
Clave de contexto	Sin	Conjuntos de datos con claves de contexto de inicio y fin de la frase.
	Con	Conjuntos de datos sin claves de contexto de inicio y fin de la frase.
Método de descuento	Lineal	Método de descuento lineal
	Witten-Bell	Método de descuento de Witten-Bell

Tabla 1. Factores del experimento

Factores de Control: Los factores y sus correspondientes niveles son los que se muestran en la tabla 1.

Factores que se deben mantener constantes: Estructura de las frases con las que se generan los archivos de prueba de los modelos de lenguaje.

Variable de Respuesta: En la tabla 2 se presenta la información relativa a la variable de respuesta considerada para este experimento.

b. Elección del diseño experimental.

Se plantea un experimento factorial 2^2, debido a que se tienen dos factores y dos niveles por cada factor. Los 20 grupos de prueba fueron seleccionados de forma aleatoria de un conjunto de frases que contienen las 20 palabras del sistema de reconocimiento.

c. Realización del experimento.

El experimento consistió en la evaluación de la perplejidad de diferentes modelos de lenguaje (uno por cada interacción de los niveles de los dos factores) para veinte grupos de prueba independientes. En la figura 5 se muestra el valor promedio de perplejidad para cada uno de los tratamientos.

d. Análisis estadístico de los datos.

Utilizando análisis de varianza multifactorial, se determinó la influencia del uso de claves de contexto y el método de descuento en la perplejidad del modelo de lenguaje. Luego, se determinó cuál tratamiento presentaba menor nivel de perplejidad a través de la prueba de comparaciones múltiples de Tukey.

Característica medida	Variable de respuesta	Rango de operación normal	Precisión de la medición.	Relación entre la variable y el objetivo
Número de posibles unidades (palabras) para la elección de una de ellas.	Perplejidad (Posibilidades por unidad)	Depende del tamaño del modelo de lenguaje.	Verificación manual	Comparación entre los valores de perplejidad obtenidos.

Tabla 2. Variable de respuesta.

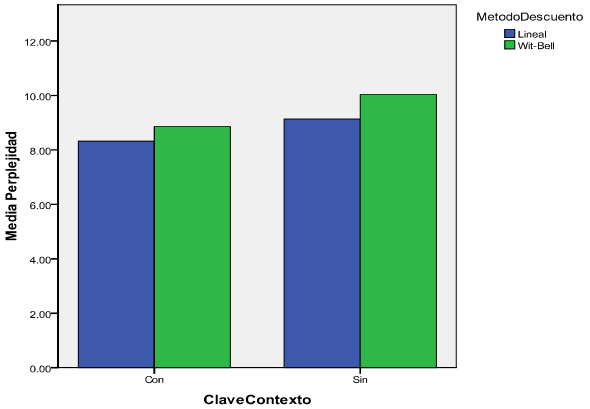

Figura 5. Perplejidad promedio para los tratamientos

e. Conclusiones y recomendaciones.

A partir del análisis experimental, puede concluirse que con un nivel de significancia de 0.05 por lo menos dos de las medias de los tratamientos son estadísticamente diferentes y que el valor de perplejidad es afectado por el uso de claves de contexto y por el método de descuento pero no por su interacción.

Luego del análisis por pares, se concluye que con un nivel de significancia de 0.05 el tratamiento con menor valor de perplejidad y por lo tanto el mejor modelo de lenguaje es el que cuenta con claves de contexto y con el método lineal como estrategia de descuento. Por esta razón, se selecciona este modelo de lenguaje para el sistema de reconocimiento de voz que se propone en este trabajo.

2) Sistema de ASR

a. Planeación previa al experimento.

Título del experimento: Estudio de técnicas de Reconocimiento Automático de Voz (ASR por sus siglas en inglés) para hablantes colombianos de la región andino-oriental.

Objetivos del experimento:

• Cuantificar la tasa de reconocimiento de sistemas de ASR para múltiples hablantes, utilizando diferentes parámetros en las técnicas de análisis de algunas de las etapas del sistema de reconocimiento. Los parámetros que se varían así como sus valores se definen como factores de control.

• Determinar la influencia de las técnicas y sus interacciones en la tasa de reconocimiento.

• Cuantificar las diferencias en el reconocimiento entre los diferentes sistemas de ASR en estudio y determinar cuál de ellos presenta un mejor nivel de reconocimiento para múltiples hablantes.

Factores de Control: Los factores y sus correspondientes niveles son los que se muestran en la tabla 3.

Factores que se deben mantener constantes: el diccionario que contiene el mapeo entre las unidades del modelo de lenguaje y las unidades del modelo acústico.

Factor	Nivel	Descripción
Dither[1] (Front-End)	Sin	Datos de entrada sin aplicación de dither
	Con	Datos de entrada con aplicación de dither
Ventana (Front-End)	Hamming	Ventana Hamming en la extracción de características
	Hanning	Ventana Hanning en la extracción de características
Modelo Acústico	CI	Unidades de voz independientes del contexto
	CD_1G	Unidades de voz dependientes del contexto con 1 gausiana por estado
	CD_2G	Unidades de voz dependientes del contexto con 2 gausianas por estado
	CD_4G	Unidades de voz dependientes del contexto con 4 gausianas por estado
	CD_8G	Unidades de voz dependientes del contexto con 8 gausianas por estado

Tabla 3. Factores del experimento

[1] Ruido aplicado a la señal de audio buscando minimizar el error de cuantización.

Característica medida	Variable de respuesta (%)	Precisión de la medición.	Relación entre la variable y el objetivo
Tasa de error de reconocimiento de palabras	- WER[2] (%)	Verificación manual	Comparación directa respecto al porcentaje de reconocimiento deseado
Tiempo de procesamiento	- TIME (s)	Verificación manual	Medición del tiempo de procesamiento requerido para reconocer cada conjunto de datos
Uso de memoria	- MEM(MB)	Verificación manual	Medición del uso de memoria requerida requerida durante el reconocimiento de cada conjunto de datos

Tabla 4. Variable de respuesta.

Factores a los que se les permite variar: El hablante es un factor que afecta el desempeño del sistema de reconocimiento de voz pero debido a que su uso por parte de diferentes hablantes es uno de los objetivos del trabajo, este factor puede y debe variar durante el experimento.

Factores perturbadores: Los siguientes son factores perturbadores no controlables identificados en el desarrollo del presente experimento:

- Ruido en el ambiente
- Estado de ánimo del hablante.
- Estado de saludo del hablante.

Variable de Respuesta:

En la tabla 4 se presenta la información relativa a las variables de respuesta consideradas para este experimento.

b. Elección del diseño experimental.

Se plantea un experimento factorial con dos factores de dos niveles y un factor de cinco niveles. El tamaño de la muestra se encuentra restringido a 352 palabras (78 frases) pronunciadas por 26 hablantes.

c. Realización del experimento.

Este sistema es evaluado por la tasa de error en el reconocimiento de palabras (WER), el tiempo de procesamiento y el uso de memoria, de donde surgen tres experimentos diferentes.

Para el desarrollo de los experimentos se seleccionaron de forma aleatoria 78 frases pronunciadas por 26 hablantes diferentes. Las frases fueron divididas de forma equitativa respecto al número de palabras en seis grupos. Cada grupo fue procesado por cada una de las propuestas para obtener valores para el WER, tiempo de procesamiento y uso de memoria. Es decir, se realizaron mediciones repetidas de los tres indicadores en los mismos grupos de datos luego de ser procesados por cada una de las propuestas de reconocimiento.

d. Análisis estadístico de los datos.

El primer experimento consistió en la evaluación de la tasa de error en el reconocimiento de palabras para diferentes propuestas de sistemas de reconocimiento automático de voz. Debido a que los datos no cumplen el supuesto de normalidad, requerido por el ANOVA, se utilizó el análisis de varianza de Friedman.

El segundo experimento consistió en la evaluación del tiempo de procesamiento durante el reconocimiento, para diferentes propuestas de sistemas de reconocimiento automático de voz. En este caso se utilizó el análisis de varianza (ANOVA) con medias repetidas. Para identificar los tratamientos con medias diferentes, se realizaron comparaciones por pares de tratamientos realizando la corrección de Bonferroni al nivel de significancia.

El tercer experimento consistió en la evaluación del uso de memoria durante el reconocimiento, para diferentes propuestas de sistemas de reconocimiento automático de voz. Se utilizó el análisis de varianza (ANOVA) con medias repetidas y para identificar los tratamientos con medias diferentes, se realizaron comparaciones por pares realizando la corrección de Bonferroni.

e. Conclusiones y recomendaciones.

A partir del análisis de los resultados del primer experimento puede concluirse que con un nivel de significancia de 0.05 no se puede afirmar que las tasas de error de reconocimiento de palabras son estadísticamente diferentes para los tratamientos.

A partir del análisis de los resultados del segundo experimento, puede concluirse que con un nivel de significancia de 0.05 se puede afirmar que el tiempo de procesamiento es estadísticamente diferente para los tratamientos. Se determina que los tratamientos sin Dither, con ventana Hamming y modelos acústicos CD_8G y CD_4G, son los que presentan un menor tiempo de procesamiento.

A partir del análisis de los resultados del tercer experimento, puede concluirse que con un nivel de significancia de 0.05 se puede afirmar que el uso de memoria es estadísticamente diferente para los tratamientos. Se determina que los tratamientos sin Dither, con ventana Hamming o Hanning y modelo acústico CD_2G, son los que presentan un menor uso de memoria.

Tratamiento			WER (%)		TIME (s)		MEM (MB)	
Dither	Ventana	Contexto	Prom	DesvEst	Prom	DesvEst	Prom	DesvEst
Con	Hamming	CD_8G	4,5638	2,83	1,2317	0,07	54,8367	1,71
Con	Hamming	CD_4G	4,5638	2,76	1,2350	0,05	52,3983	1,25
Con	Hamming	CD_2G	4,2715	4,61	1,2917	0,06	50,4250	2,63
Con	Hamming	CD_1G	4,8267	4,2	1,4117	0,11	51,7700	2,32
Con	Hamming	CI	4,8218	2,83	1,8917	0,07	51,3617	1,82
Con	Hanning	CD_8G	4,8413	1,82	1,2467	0,05	55,3500	1,61
Con	Hanning	CD_4G	4,2715	4,6	1,2467	0,08	52,8517	3,01
Con	Hanning	CD_2G	4,8365	5	1,2850	0,07	50,6383	4,35
Con	Hanning	CD_1G	5,1097	1,82	1,4283	0,07	52,1500	2,9
Con	Hanning	CI	3,9743	2,01	1,9150	0,07	53,7467	1,74
Sin	Hamming	CD_8G	7,1062	4,97	1,2250	0,05	53,4550	3,44
Sin	Hamming	CD_4G	6,5313	4,97	1,1867	0,08	51,6633	3,2
Sin	Hamming	CD_2G	6,5313	2,25	1,2467	0,05	49,5950	1,56
Sin	Hamming	CD_1G	5,6692	1,05	1,3650	0,06	53,6483	1,26
Sin	Hamming	CI	4,2565	4,98	1,8633	0,08	51,5483	3,04
Sin	Hanning	CD_8G	7,0963	4,76	1,2417	0,08	53,4133	1,78
Sin	Hanning	CD_4G	5,9663	2,71	1,2117	0,09	53,8383	2,05
Sin	Hanning	CD_2G	6,5313	1,39	1,2550	0,07	49,6283	0,9
Sin	Hanning	CD_1G	5,9517	4,88	1,4117	0,13	51,4117	2,72
Sin	Hanning	CI	4,8217	4,72	1,8883	0,11	51,3783	2,6

Tabla 5. Promedios de los tratamientos

[2] WER = 100 * No. De palabras no reconocidas / No. De palabras pronunciadas.

En la tabla 5 se muestran los promedios y desviación estándar de la tasa de error de reconocimiento de palabras, tiempo de procesamiento y uso de memoria para cada uno de los tratamientos. En el caso del tiempo de procesamiento y el uso de memoria se resaltan en azul los promedios de los mejores tratamientos. En el caso de la tasa de error en reconocimiento, el análisis experimental determinó que no hay diferencia significativa entre los promedios de los tratamientos.

Los resultados obtenidos en tasa de reconocimiento de palabras de la tabla 5 son superiores a los encontrados en la literatura para estudios con características similares. En [28] se presenta el uso de Kernels y HMM discretos para el reconocimiento de dígitos en español. El sistema se entrenó con 396 pronunciaciones de dígitos y se validó con 197 pronunciaciones realizadas por 6 adultos. El sistema con HMM presentó un WER de 9.65% y el sistema con Kernels uno de 8.63%.

En [29] se propone un sistema de reconocimiento de comandos en español usando mezclas de Gaussianas. Los datos de entrenamiento estaban compuestos por 40 grabaciones de los comandos y los de validación por 20 grabaciones. El sistema presentó un WER de 5.6%, que es un valor comparable con los resultados obtenidos para las propuestas presentadas en la tabla 2.

V. CONCLUSIONES

En este artículo se presentó una propuesta metodológica para construir un sistema de reconocimiento automático de voz basado en Modelos Ocultos de Markov utilizando Sphinx-4 como herramienta de implementación y técnicas de diseño de experimentos y análisis de datos para la selección del mejor sistema de reconocimiento. Los indicadores de desempeño del sistema utilizados fueron el WER, el tiempo de procesamiento y el uso de memoria durante el reconocimiento.

La propuesta de ASR fue implementada en Sphinx-4.0. El sistema se encuentra orientado a hablantes andino-Orientales de nacionalidad Colombiana. El vocabulario se enmarca en un sistema de marcado automático y se limita a un conjunto de veinte palabras.

Los resultados obtenidos muestran un funcionamiento satisfactorio en la tasa de reconocimiento que es superior o comparable con otros estudios similares para lengua castellana.

REFERENCIAS

[1] C. Llamas y V. Cardeñoso, Reconocimiento automático del habla, técnicas y aplicación, Universidad de Valladolid, 1997.

[2] X. Zhao and J. Wang, "A new noisy speech recognition method", IEEE International Symposium on Communications and Information Technology, Volume 1, p. 292 – 296, 12-14 Oct. 2005.

[3] L. Rabiner and B. Juang, Fundamentals of speech recognition, Prentice Hall, New Jersey 1993.

[4] S. Tsuge, M. Shishibori, K. Kita, F. Ren and S. Kuroiwa, "Study of Intra-Speaker's Speech Variability over Long and Short Time Periods for Speech Recognition", Proc. of 2006 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP2006), Vol.1, p. 397-400, Toulouse, France, May 2006.

[5] M. Nazmi, Q. Zhi, A. Cheok, K. Sengupta, J. Zhang and C. Chung. "Analysis of lip geometric features for audio-visual speech recognition", IEEE Transactions on Systems, Man, and Cybernetics, p. 564-570 (2004).

[6] M. Akbacak and J. Hansen, " Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems", , IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, p. 465 – 477, Feb. 2007.

[7] D. Montgomery, Diseño y Análisis de experimentos, 2nd ed., Limusa Wiley , Mexico 2007.

[8] F. Casacuberta y E. Vidal, Reconocimiento automático del habla, Marcombo Boixareu editores, 1987.

[9] G. Santos, Inteligencia artificial y matemática aplicada: reconocimiento automático del habla, Universidad de Valladolid, 2001.

[10] J. Deller and J. Proakis, J. Hansen, Discrete-Time processing of speech Signals, Macmillan Publishing Company, New York, 1993.

[11] B. Rapp, "N-gram language models for polish language. basic concepts and applications in automatic speech recognition systems", Proceedings of the International Multiconference on Computer Science and Information Technology, p. 321-324, 2008.

[12] T. Kubik and M. Sugisaka, "Use of a cellular phone in mobile robot voice control", Proceedings of the SICE Nagoya, p. 106–111, July 25–27 (2001).

[13] M. Benzeguiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. issore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi and C. Wellekens, "Automatic Speech Recognition and Intrinsic Speech Variation", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 5, 14-19 May 2006.

[14] M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer", IEEE Transactions on acoustics, speech and signal processing, p. 400-401, 1987

[15] Coleman D, Montgomery D, "A Systematic Approach to Planning for a Designed Industrial Experiment", on Technometrics. Publication Date: February 1993, Vol 35 No.1.

[16] R. Young y D. Veldman, Introducción a la estadística aplicada a las ciencias de la conducta, Editorial Trillas S.A., México 2005.

[17] CMU Sphinx, Open source toolkit for speech recognition, Project bu Carnegie Mellon University. http://cmusphinx.sourceforge.net/wiki/

[18] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition".

[19] J. Montes, El español de Colombia. Propuesta de clasificación dialectal, THESAURUS Tomo XXXVII. Num 1 (1982).

[20] Z. Jiang, H. Huang, S. Yang and Z. Hao, "Acoustic Feature Comparison of MFCC and CZT-Based Cepstrum for Speech Recognition", Fifth International Conference on Natural Computation, p. 55-59. Aug. 2009.

[21] J. Vojtko, J. Korosi and G. Rozinaj, "Comparison of automatic speech recognizer SPHINX 3.6 and SPHINX 4.0 for creating systems in Slovak language", 15th International Conference on Systems, Signals and Image Processing, p. 537-539. June 2008.

[22] T. Kambe, H. Matsuno, Y. Miyazaki and A. Yamada, "C-based design of a real time speech recognition system", 2006 IEEE International Symposium on Circuits and Systems, p. 1755-1760. Sept. 2006.

[23] M.J.F. Gales, "Acoustic modeling for speech recognition: Hidden Markov models and beyond", IEEE Workshop on Aotmatic Speech Recognition & Understanding, p. 44-44, January 2010.

[24] A. Mantilla, H. Pérez, D. Mata, C. Angeles, J. Alvarado and L. Cabrera, "Recognition of Vowel segments in Spanish esophageal speech using hidden Markov models", 15th International conference on computing, p. 115-120, Dec 2006.

[25] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the cmu-cambridge toolkit", Proceedings of the ESCA Eurospeech Conference.

[26] J. Segura, A. Rubio, P. García y M. Benítez, "Resultados Preliminares Sobre Modelos MVQHMM". SEPLN. Boletín no. 13, p. 411-418. Febrero 1993.

[27] J. Hieronymus, ASCII Phonetics Symbols for the World´s Languages: Worldbet, AT&T Bell Laboratories, Murray Hill, NJ 07974, USA.

[28] J. Goddard, A. Martínez, F. Martínez and H. Rufiner, "A comparison of String Kernels and discrete Hidden Markov Models on a Spanish Digit Recognition Task", 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, p. 2962-2965, Cancun, Mexico, Sep. 2003.

[29] P. Mayorga, J. Olguín, A. Hernandez and J. Flores, "Gaussian Components Optimization for a Robot Controlled by Speech Commands in Mexican Spanish", Electronics, Robotics and Automotive Mechanics Conference, p. 324-329. October 2007.