DISEÑO DE PROTOTIPO PARA MEJORAR LA DICCIÓN MEDIANTE EL USO DE MODELOS OCULTOS DE MARKOV

Ángel David Pedroza Ramírez

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica

P.A.D_16@hotmail.com

José Ismael de la Rosa Vargas

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica

joseismaelrv@gmail.com, ismaelrv@ieee.org

Ernesto García Domínguez

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica

ers680807@yahoo.com.mx

Hamurabi Gamboa Rosales

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica

hamurabigr@hotmail.com

Aldonso Becerra Sánchez

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica

A7donso@hotmail.com

Resumen

La comunicación oral en el ser humano es muy importante, sin embargo, la buena comunicación, independientemente del idioma, debe ser clara, objetiva y expresiva con el fin de que lo que se quiere expresar sea lo que el oyente entienda. El reconocimiento de voz, por otro lado, se basa en el estudio sobre el proceso del habla y la comunicación, y la forma en que este conocimiento puede ser aplicado como herramienta para diversas finalidades. El enfoque de esta

investigación es el desarrollo de un prototipo didáctico para realizar pruebas de dicción en el idioma español. Para ello, se utilizaron 3 técnicas basadas en Modelos Ocultos de Markov (HMM) las cuales son Modelos Ocultos de Markov con DTW (MDTW), Modelos Ocultos de Markov con DTW aproximado por izquierda y derecha (MID) y Modelos Ocultos de Markov con relleno de palabras (MRP). Con esta estructura se logró distinguir entre calidades de dicción y con una eficiencia de reconocimiento por encima del 90 % para cualquiera de las técnicas utilizadas. Finalmente, con base en lo anterior, se programó una interfaz en Matlab la cual brinda resultados para la corrección de la dicción.

Palabra(s) Clave(s): Dicción, GUI, Interfaz didáctica, Reconocimiento Automático de Voz.

1. Introducción

El ser humano ha desarrollado la capacidad de comunicación dada su necesidad de vivir en sociedad. Gracias a la generación de sonidos articulados, la comunicación oral substituyó a la mayoría de las técnicas primitivas. Sin embargo, la buena comunicación, independientemente del idioma, debe ser clara, objetiva y expresiva con el fin de que lo que se quiere expresar sea lo que el oyente entienda [6].

Hace algunas décadas se pensaba que educar la voz era asunto de los comentaristas, locutores, reporteros, entre otros. Hoy en día, sin embargo, no se puede dejar este asunto para aquellas áreas antes mencionadas sino que cualquier persona necesita hablar un lenguaje correcto, claro y natural.

Por otro lado, la información que se transmite a través de la voz posee características que pueden ser analizadas por medio de sistemas enfocados en reconocimiento automático de voz [1, 2, 3, 4]. El reconocimiento de voz es una rama de la investigación multidisciplinaria que ha logrado dar solución a diversas problemáticas del mundo real (desde aplicaciones en telefonía, hasta sistemas de ayuda para personas con discapacidades físicas o en rehabilitación [5]). Específicamente el reconocimiento de voz es de nuestro especial interés dado que mediante este podemos lograr, una vez extraída la información por algún método,

reconocer palabras con algún margen de error para después utilizarla en algún proceso.

La señal de voz es una señal de tipo aleatoria por lo que es necesario utilizar algún método que sea capaz de tomar en cuenta esta característica. Los Modelos Ocultos de Markov (Hidden Markov Models - HMM) son una de las herramientas estocásticas que han demostrado tener un buen desempeño en tareas de reconocimiento de voz. En este sentido, para lograr el reconocimiento de una palabra en específico, es necesario contar con repeticiones de la misma palabra (corpus de voces) y con esa información, mediante Modelos Ocultos de Markov, crear un modelo específico para dicha palabra. Sin embargo, es necesario tomar en cuenta que en las repeticiones de voz existen variaciones inter e intra-locutor aunado a otros factores (duración, velocidad, etc.). Dado que la información que utilizan los Modelos Ocultos de Markov son esas repeticiones, es necesario tomar en cuenta que estas influyen directamente en la calidad del modelo desarrollado para cada palabra a reconocer.

Esta investigación plantea el desarrollo de un prototipo didáctico para realizar pruebas de dicción de palabras aisladas en el idioma español. Para ello se estableció el vocabulario a identificar, la creación de un corpus de voz adecuado para el idioma español, el pre-procesamiento y extracción de características (tomando para este último a la energía como parámetro). Además se propusieron 3 esquemas diferentes de reconocimiento basados en Modelos Ocultos de Markov las cuales son Modelos Ocultos de Markov con DTW (MDTW), Modelos Ocultos de Markov con DTW aproximado por izquierda y derecha (MID), y Modelos Ocultos de Markov con relleno de palabras (MRP). Con la estructura de los algoritmos presentados, se logró distinguir entre calidades de dicción contando con una eficiencia de reconocimiento por encima del 90 % (para cualquiera de las técnicas utilizadas). Finalmente se programó una interfaz en Matlab la cual brinda resultados para la corrección de la dicción.

El artículo se organiza de la siguiente manera. La sección 2 se enfoca en algunos fundamentos sobre la dicción y las unidades fonéticas. En la sección 3 se presentan las técnicas de reconocimiento de voz fundamento para esta

investigación (HMM y DTW). En la sección 4 se explican brevemente los métodos propuestos. La sección 5 presenta una explicación del prototipo desarrollado. Finalmente en la sección 6, se presentan algunas conclusiones.

2. Estructura del habla

Fonemas

El conjunto de fonemas (cada uno de los sonidos simples del lenguaje hablado) en combinaciones diferentes, forman las palabras propias de un idioma y la base de la comunicación oral. De tal forma que cada idioma posee un conjunto propio de fonemas (por ejemplo, para el español 24, para el inglés cerca de 42 y para el francés casi 50 fonemas). Para el caso del idioma español la subdivisión de los fonemas es la siguiente [6]:

- **Timbres Básicos**: son variaciones del sonido producido naturalmente por las cuerdas vocales. Para el caso del idioma español: /a/, /e/, /i/, /o/ y /u/.
- La correcta producción de dichos fonemas (sin esfuerzo y con naturalidad) es el material de estudio en la impostación de la voz.
- **Sonidos auxiliares**: Se producen por el aparato resonador sin intervención de las cuerdas vocales. Corresponden a las consonantes sonoras: /m/, /n/, /l/, /s/, /j/, /r/.
- **Ataques**: Son modos de iniciar un timbre básico o un sonido auxiliar; representan las distintas formas de la oclusión que impide al aire fluir. Según la brusquedad de la liberación de aire, los ataques se dividen en:
    - ✓ Fuertes: /t/, /k/.
    - ✓ Suaves: /b/, /p/, /d/.
    - ✓ Mixtos: Como caso particular de ataques, los sonidos mixtos, se refiere a la sucesión de sonidos auxiliares /l/ o /n/ y el timbre básico /i/ (/ll/ y "/ñ/"); o las combinaciones /k/ con /s/ (/x/).

Es necesario conocer las reglas de cada idioma para formar combinaciones entre fonemas.

Dicción y fonación

La dicción, como tema central de la investigación, se refiere a la correcta producción de los fonemas y la combinación con los timbres básicos. En otras palabras, es la forma de emplear palabras de forma correcta y acertada en el idioma al que pertenecen sin poner especial interés en lo que estas quieren expresar. Para lograr tener una dicción excelente es fundamental poseer una pronunciación correcta y matizar los sonidos respetando las pausas que sean necesarias.

La dicción, buena o mala, no tiene que ver con el significado que se desea transmitir ni con aquello que se pretende expresar. Cuando se posee una manera clara de hablar y resulta fácil entender a quien se expresa, se habla de una dicción clara o limpia.

Contrario a la buena dicción se encuentran los vicios del lenguaje o defectos en el habla. Sobre este punto, es necesario aclarar que la dicción no está precisamente vinculada al entendimiento; algunas palabras que no poseen buena dicción aun así son reconocibles pero no es tan sencillo de entender como aquellas que poseen buena dicción. Ejemplo de ellas son el utilizar "veniste" en vez de "viniste"; o "andabanos" en vez de "andábamos".

Se puede lograr una dicción correcta y por ende un habla clara, pronunciando cada palabra y sílaba de forma adecuada con el fin de incrementar la memoria muscular en base a repeticiones. Para mejorar la dicción es necesario también tomar en cuenta la velocidad a la que son emitidas las palabras por lo que, se recomienda una velocidad pausada con el fin de prestar atención a cada parte de la estructura del mensaje oral que se desea transmitir; la rapidez excesiva puede generar problemas en la comunicación. Esto no debe ser un sinónimo de monotonía y aburrimiento sino una mejoría en la voz y el ritmo.

Para poder poseer una buena fonación es indispensable poseer las siguientes características [6]:

- Suficiente: Mediante el dominio de la respiración.
- Clara: Correcta producción de cada uno de los sonidos que forman el idioma (aislada o en combinaciones); es decir, con dicción.

- Expresiva: En entonación, ritmo, intensidad y timbrado para poder captar no solo el significado propio de la palabra pronunciada sino el significado explícito dado por la entonación, velocidad y pausas.

Existen dos ramas de estudio muy importantes dentro de la teoría de producción de voz: Impostación y Dicción. En lo que se refiere al término conocido como "impostación de la voz" se refiere al aprovechamiento de la espiración (expulsión de aire por los pulmones) para producir sonido con el máximo rendimiento y el minino esfuerzo en la garganta.

Tal como lo indica la teoría, se necesita que el aire expulsado por los pulmones sea totalmente puesto en vibración por las cuerdas vocales; y éstas a su vez vibren por la acción del aire y la tensión propia que poseen; y una vez cumplidas estas condiciones, el aire se aproveche en el aparato resonador de forma conveniente. Sin embargo, algunos de los hábitos adquiridos hacen que la capacidad fonadora presente obstáculos en algunas de las etapas descritas.

En conclusión, para poseer una buena dicción es necesario pronunciar correctamente, acentuar adecuadamente y poseer buena velocidad. Es por ello que, es necesario el desarrollo de un método capaz de tomar en cuenta estas tres características y con ello determinar si se tiene una excelente, buena o mala dicción.

3. Técnicas de reconocimiento de voz

Para la presente investigación se utilizaron las siguientes técnicas de reconocimiento:

Modelos Ocultos de Markov (HMM)*:* La aproximación de patrones no modela estadísticamente a las señales de voz por lo que posee ciertas restricciones. Por otro lado, las técnicas estadísticas han sido aplicadas en el agrupamiento para crear patrones de referencia. Mediante ello se mejora la clasificación y se realiza una simplificación dado que se utilizan múltiples señales de referencia y se caracterizan mejor las variaciones de diferentes pronunciaciones*.* Una de las herramientas que permiten caracterizar estadísticamente las propiedades de un

patrón o señal de voz son los Modelos Ocultos de Markov (MOM ó HMM). Un modelo oculto de Markov está caracterizado por [7,8]:

- El número de estados en un modelo (N).
- El número de símbolos distintos de observación en cada estado (M).
- La matriz de probabilidad de transición de estados ($A = [a_{ij}]$).
- La distribución de probabilidad de los símbolos en el estado ($B = [b_j(k)]$).
- La distribución de probabilidad inicial de estados ($\pi = [\pi_i]$).

De manera simplificada un HMM puede ser expresado en ecuación 1.

$$\lambda = (A, B, \pi) \quad (1)$$

En la ecuación 1, λ es el nombre de un modelo particular o palabra específica a identificar.

En la aplicación de un HMM es necesario asumir que la señal de voz es en su naturaleza una señal aleatoria y cuyos parámetros pueden ser estimados. Es por ello que los HMM son una herramienta de gran utilidad para tareas diversas en reconocimiento de voz. Para aplicar un HMM son necesarias dos etapas:

- Entrenamiento: Creación de los modelos por palabra a reconocer.
- Prueba: Reconocimiento de la palabra con los modelos previamente entrenados. Es decir, obtención de la probabilidad de que la repetición a identificar haya sido generada por un modelo de palabra entrenada específico.

Alineamiento Dinámico en el Tiempo (DTW)*:* Debido a que las palabras a reconocer, y por tanto la información extraída, posee variaciones (escalares, temporales, entre otras), la comparación entre muestras resulta un tanto complicada. Es por ello que es de gran utilidad realizar un alineamiento acústico temporal para disminuir estas variaciones.

El proceso anterior se puede llevar a cabo mediante el Alineamiento Dinámico en el Tiempo (DTW) [9]. Este método busca un empate entre muestras de dos diferentes repeticiones vocales de la misma palabra (buscando causar el mínimo desajuste en las repeticiones).

Lo que se busca es por tanto, una ruta óptima de empate entre muestras. El "algoritmo de programación dinámica" permite obtener la ruta óptima sin tener que obtener todas las rutas posibles. El alineamiento se realiza minimizando localmente la distancia entre muestras. La distancia acumulada se obtiene mediante ecuaciones 2 y 3.

$$D(n,m) = d(n,m) + \min[D(n-1,m)g(n-1,m), D(n-1,m-1), D(n-1,m-2)] \quad (2)$$

$$g(n,m) = \begin{cases} 1, \text{para } w(n) \neq w(n-1) \\ \infty, \text{para } w(n) = w(n-1) \end{cases} \quad (3)$$

En donde d(n, m) es la distorsión en tiempo y w (.) es la función de alineamiento entre n y m muestras.

4. Métodos propuestos

Como se describió anteriormente, los Modelos Ocultos de Markov son de gran utilidad cuando se trabaja con información estocástica (como la contenida cuando se analiza la señal de voz). Una vez que se han analizado algunas de las metodologías comúnmente utilizadas para el reconocimiento de voz, de entre dicha variabilidad, se proponen varios algoritmos enfocados en la realización de pruebas de dicción.

Dada la capacidad que poseen los HMM para trabajar con estados, es que se pretende buscar un método capaz de generar dichos estados para ser posteriormente tratados. Sin embargo, si las repeticiones de voz (que producen dicho los estados) con las que trabajan los HMM varían en cuanto a velocidad y duración, es necesario utilizar una técnica que pueda disminuir el efecto que producen estas variaciones. Tomando en cuenta lo anterior, para la investigación presente, se utilizaron Modelos Ocultos de Markov junto con algunas modificaciones las cuales son: Modelos Ocultos de Markov con DTW, Modelos Ocultos de Markov con DTW aproximado por izquierda y derecha y Modelos Ocultos de Markov con relleno de palabras.

Modelos Ocultos de Markov con DTW (MDTW)

Basado en que algunas investigaciones recientes que han utilizado mezclas entre DTW y HMM han mostrado buenos porcentajes en tareas de reconocimiento de voz, [se propone alinear temporalmente los eventos acústicos (repeticiones) antes de crear los modelos por palabras [10]. El diagrama de flujo de este algoritmo se muestra en la figura 1.

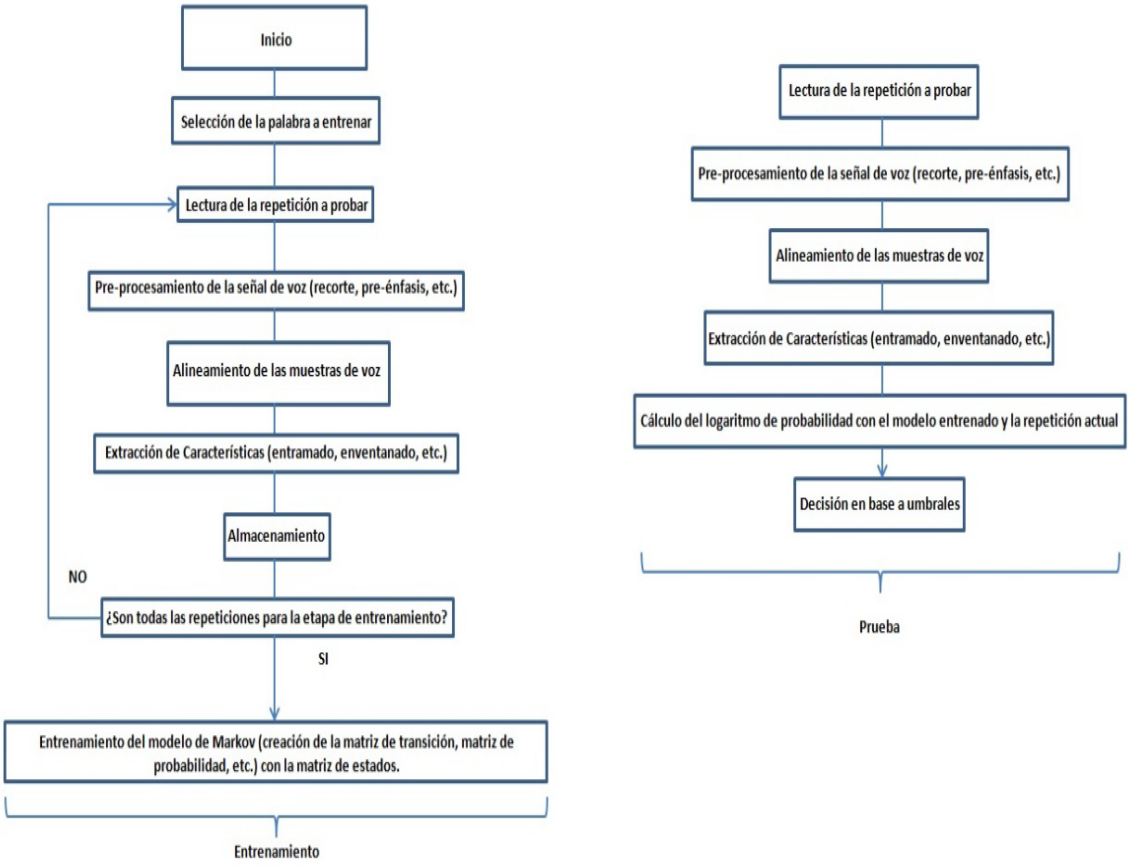

Figura 1 Diagrama de flujo para MDTW.

Este proceso de alineamiento se explica como sigue: Una vez que la señal es tratada por la fase de procesamiento (mediante el recorte, pre-énfasis, segmentación, etc.) la señal de voz $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}$ posee variabilidad respecto de la señal $\mathbf{Y} = \mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_M}$, la cual se presupone posee “buena dicción” y respecto de la cual se desea hacer el alineamiento. Debido a que generalmente los subíndices *N* y *M* no son iguales, es necesario encontrar una función que

relacione la distorsión entre muestras. Esta función se expresa en ecuación 4.

$$D(x,y) = \sum_{n=1}^{N} d(n,m) \quad (4)$$

Una vez se encuentra la función de distorsión y se realiza el Alineamiento Dinámico, se procede (como se muestra en el diagrama de flujo de la figura 1) a la extracción de características y se continúa con el proceso de entrenamiento y prueba.

Este alineamiento permite que las matrices de características con las que se entrenan los HMM sean cuadradas (mismo número de características para todas las señales en entrenamiento de la misma palabra).

Modelos Ocultos de Markov con DTW aproximado por izquierda y derecha (MID)

Algunas de las palabras del vocabulario a reconocer pueden poseer similitudes fonéticas (como en el caso entre las palabras “veinti-uno” y “veinti-dos”, o entre “carro-za y “carro-cería” por poner algunos ejemplos). Tomando en cuenta estas similitudes, el método propuesto considera que la señal de voz puede ser dividida en dos sub-unidades fonéticas (como el caso utilizado en domótica para el reconocimiento de palabras continuas [11]).

El método de Modelos Ocultos de Markov con DTW aproximado por izquierda y derecha (MID) consiste en, procesada y alineada la señal de voz $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}$ como en el método anterior, dividir la señal en dos sub-unidades fonéticas y tratar a cada una por separado para las etapas de entrenamiento. Esto resulta en la creación de dos modelos por palabra bajo entrenamiento, ecuaciones 5 y 6.

$$\lambda 1 = (A, B, \pi) \dots\dots\dots \text{Modelo parte izquierda} \quad (5)$$

$$\lambda 2 = (A, B, \pi) \dots\dots\dots \text{Modelo parte derecha} \quad (6)$$

Una vez se almacenan los modelos entrenados, a la señal bajo identificación $\mathbf{Z} = \mathbf{z_1}, \mathbf{z_2}, \dots, \mathbf{z_N}$, se le aplica el mismo proceso de Alineamiento Dinámico y división en 2 unidades fonéticas. Finalmente, el reconocimiento se lleva a cabo obteniendo la probabilidad del modelo entrenado de la parte izquierda y derecha de la palabra

entrenada respecto de la parte izquierda y derecha de la señal bajo reconocimiento. El diagrama de flujo de este algoritmo se muestra en la figura 2.

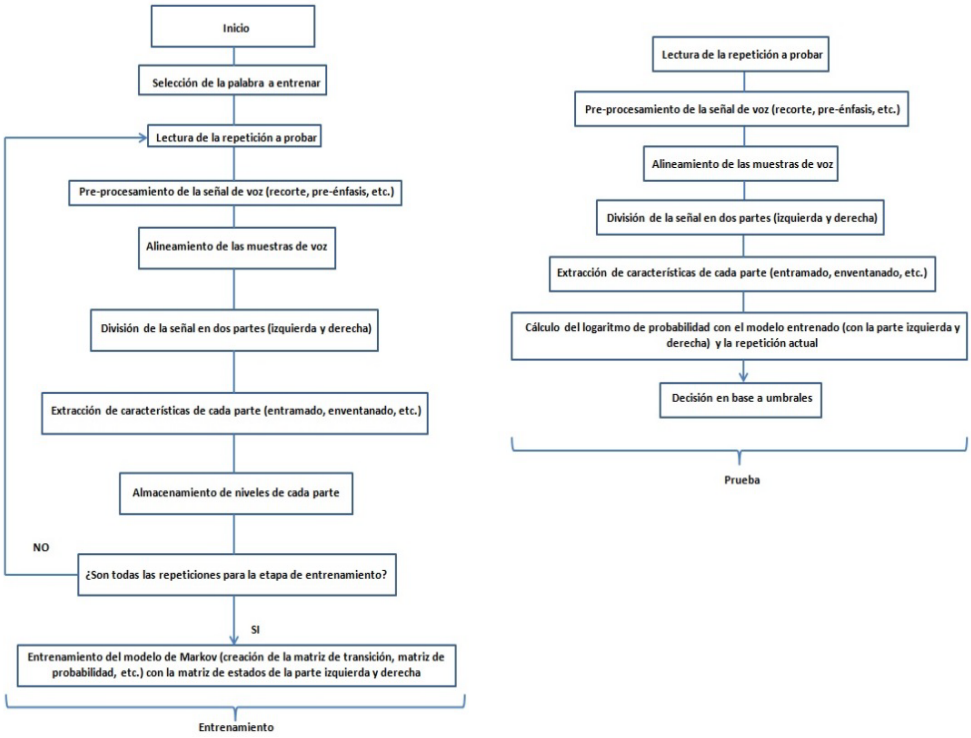

Figura 2 Diagrama de flujo para MID.

Modelos Ocultos de Markov con relleno de palabras (MRP)

El Alineamiento Dinámico en el tiempo (DTW), en mayor o menor medida, modifica la información acústica presente en cada repetición alineada. Una forma de disminuir este efecto y asegurar el mismo número de muestras para todas las señales en entrenamiento es crear la matriz de características ajustando cada repetición al tamaño más grande de las muestras.

Supongamos que se tiene almacenado un vector $\mathbf{V} = \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_N}$ que corresponde al bloque de N parámetros de características de la palabra “ferrocarril”. Al extraer las características de otra de las repeticiones de la misma palabra se obtiene un vector característico $\mathbf{B} = \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_M}$ que corresponde a M parámetros de características de la palabra “ferrocarril”.

En esta situación se presentan 3 casos diferentes:

- Cuando *M* es más pequeña que *N*, se "rellena" el vector característico $\mathbf{B} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M$ al tamaño de $\mathbf{V} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$.
- Cuando M es más grande que N, se "rellena" el(los) vector(es) previamente almacenado de tamaño N al tamaño de $\mathbf{B} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M$.
- En el caso particular (ideal) cuando las muestras M y N sean del mismo tamaño, entonces el vector $\mathbf{B} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M$ se almacena tal cual sin sufrir ningún ajuste. El diagrama de flujo de este algoritmo se muestra en la figura 3.

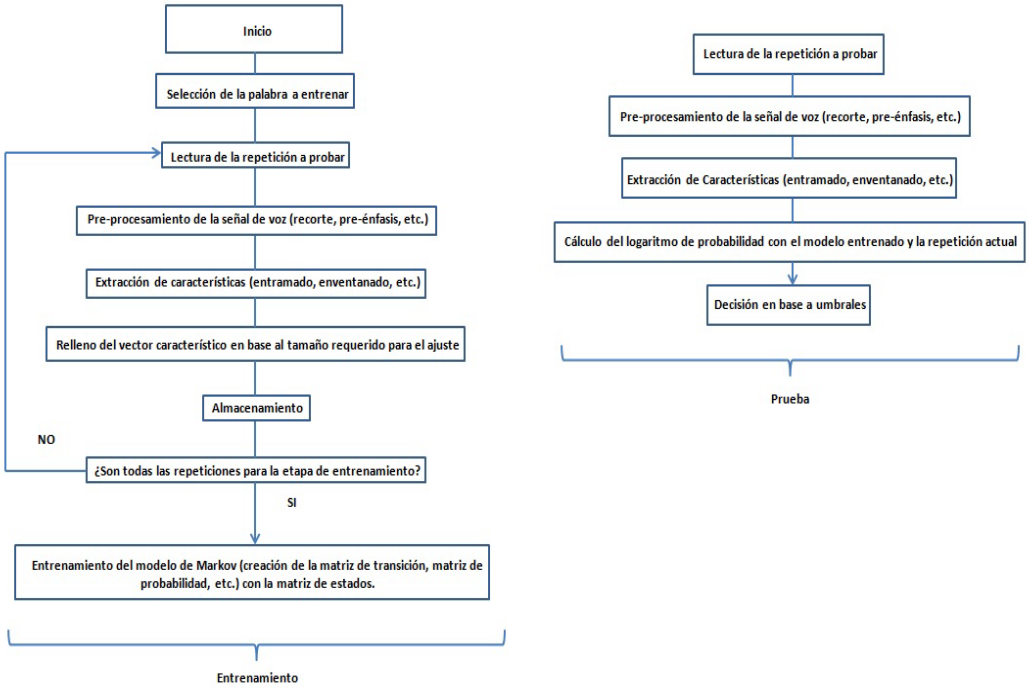

Figura 3 Diagrama de flujo para MRP.

5. Prototipo desarrollado

Una vez que se cuenta con los algoritmos, es necesario determinar la lista de palabras a identificar, la realización de pruebas con cada una de ellas y, con base en ello, desarrollar una interfaz apropiada para la realización de pruebas de dicción. En la tabla 1, se indican la lista de palabras seleccionadas y el método que mejor resultados brindó para cada una de ellas (palabras que según una

encuesta, son algunas de las más difíciles de pronunciar en el idioma español [12]).

Tabla 1 Métodos de identificación por palabras.

Palabra	**Método**
Helicóptero	MDTW
Ineptitud	MRP
Madrid	MRP
Prejuicios	MDTW
Albóndiga	MRP
Indecisión	MID

Es así que, según la palabra a identificar, es el método que la interfaz utiliza para su reconocimiento. Dado que el objetivo es el desarrollo de una interfaz prototipo para probar la dicción de un hablante común, es necesario omitir parámetros como la elección de método de reconocimiento, gráficas de información frecuencial o algunos otros que, para el hablante común, no son de mucha utilidad. Debido a que los algoritmos de reconocimiento desarrollados para la presente investigación se programaron en MATLAB 2011, se optó por desarrollar el prototipo en el entorno gráfico (GUI) de MATLAB. Tanto para la etapa de prueba como de entrenamiento de Modelos Ocultos de Markov, se utilizó una herramienta desarrollada por Kevin Murphy llamada Hidden Markov Model (HMM) Toolbox, la cual es una librería especial para Matlab que se puede encontrar de forma gratuita en línea [13].

Es por ello que, para poder utilizar la interfaz desarrollada, es necesario pre-cargar algunos archivos especiales (descargable del sitio web). La lista de archivos necesarios son HMM, KPMstats, KPMtools y netlab3.3. La estructura de la interfaz desarrollada mediante el GUIDE se describe en secciones siguientes, ejemplificando su uso.

Menú principal

En la figura 4 se muestra la interfaz inicial del programa desarrollado. En ella se tiene como elementos los botones:

- Iniciar: Abre la interfaz para la realización de nuevas pruebas de dicción.

- Ejemplos: Abre una interfaz con algunas dicciones de prueba.
- Salir.

Figura 4 Menú principal.

Ejemplos

Si se desea aprender cómo funciona la interfaz para las pruebas de dicción, es necesario presionar el botón con la etiqueta “Ejemplos” (ver figura 5).

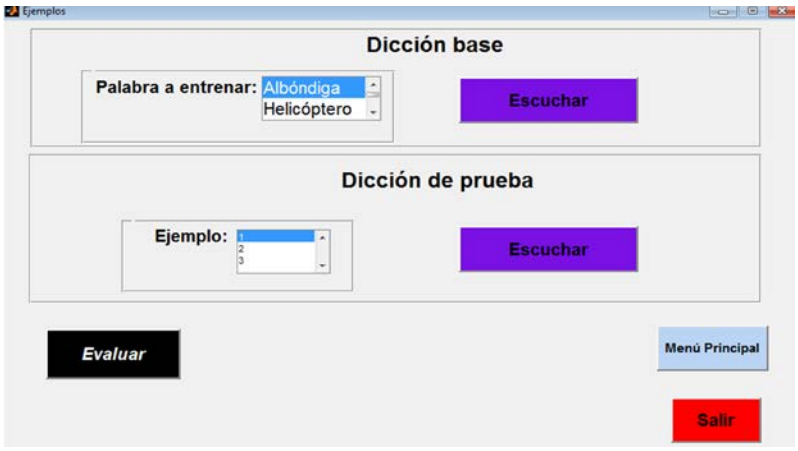

Figura 5 Sub menú “ejemplos”.

Dado que lo primero que se necesita para probar la dicción es conocer la dicción “esperada” (correcta), es necesario definir, de entre la lista de palabras posibles (vocabulario predefinido), aquella que se desea probar. Esto se lleva a cabo seleccionando de la lista “Palabra a entrenar” la palabra elegida y presionando el botón “Escuchar” de la sección “Dicción base”.

Algo a destacar es que estas repeticiones de palabras base (repeticiones con dicción correcta) fueron adquiridas (tanto para la prueba de los algoritmos como para el desarrollo de la estructura gráfica) de hablantes estudiantes de la licenciatura en comunicaciones de la Universidad Autónoma de Baja California, cuya dicción puede ser clasificada como "correcta" gracias a su preparación profesional. Dichas grabaciones fueron obtenidas en una cabina especial para video conferencias del Centro de Investigación y Desarrollo de Tecnología Digital (CITEDI - IPN).

Una vez conocido que se conoce la dicción "correcta", en la sección "Dicción de prueba" se pueden probar algunas dicciones ejemplo (3 por cada palabra) y conocer el resultado obtenido en base a la dicción encontrada. Para llevar a cabo este proceso es necesario elegirlos de la lista "Ejemplo" y presionar el botón "Escuchar". Para conocer la calificación presente en el ejemplo de la palabra seleccionada, es necesario presionar el botón "Evaluar". En la figura 6 se muestra un ejemplo del resultado obtenido.

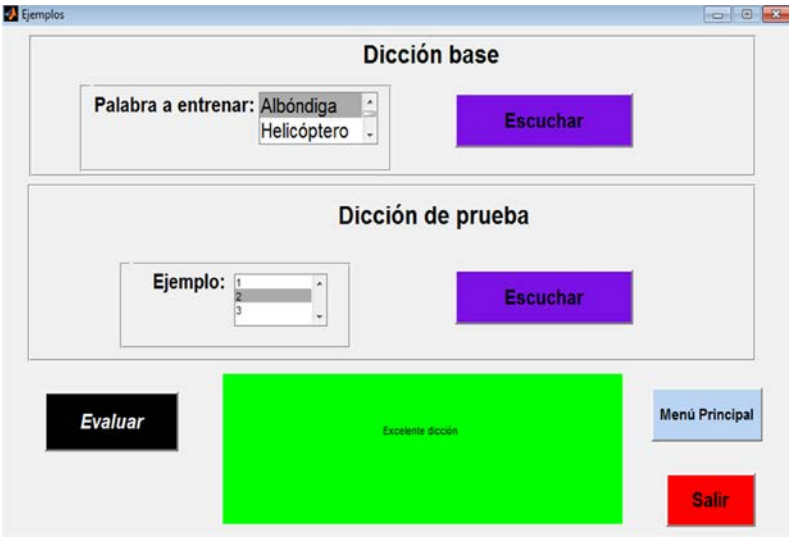

Figura 6 Ejemplo propuesto de la prueba de dicción de la palabra "albóndiga".

El proceso de calificación de dicción se programó con base al despliegue de 3 colores diferentes:

- Color verde: Excelente dicción,
- Color Amarillo: Buena dicción,

- Color Rojo: Mala dicción.

Un aspecto fundamental a destacar es que esta relación de colores es proporcional al reconocimiento realizado mediante los algoritmos propuestos. Esto se debe a que, los valores obtenidos en el reconocimiento (ver tabla 2) son valores probabilísticos los cuales, es necesario relacionar con estructuras fáciles de entender para cualquier persona que desea probar su dicción.

Tabla 2 Decisión con base a probabilidad y umbrales.

Técnica utilizada	Excelente dicción	Buena dicción	Mala dicción
MID	Loglik <100	Loglik entre 100 y 150	Loglik >150
MRP	Loglik <150	Loglik entre 150 y 300	Loglik >300
MDTW	Loglik <150	Loglik entre 150 y 300	Loglik >300

Iniciar

En la figura 7, el botón "Iniciar" del menú principal, es utilizado para probar nuevas repeticiones de voz. El proceso para seleccionar y escuchar la dicción base es similar a la descrita en la fase de pruebas.

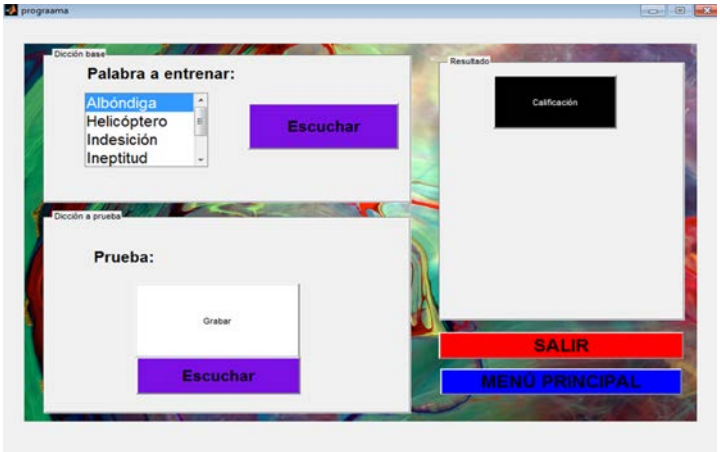

Figura 7 Sub menú "iniciar".

A diferencia de la interfaz desarrollada mostrada en figuras 5 y 7 se puede grabar y escuchar (mediante el botón correspondiente) la repetición de la palabra a probar con la finalidad de corregirla (si es que la grabación a probar no fue acústicamente satisfactoria).

Finalmente, al igual que el submenú de “opciones”, mediante el botón “Calificación”, se obtiene el resultado de la dicción encontrada.

6. Conclusiones

El prototipo presentado tiene la finalidad de ser aplicado para la realización de pruebas de dicción en el idioma español. Debido a este enfoque es que se utilizaron diversas técnicas y parámetros para calificar, de un vocabulario definido, repeticiones de voz con diferentes dicciones.

Específicamente para esta investigación, se utilizaron Modelos Ocultos de Markov junto con algunas modificaciones (Modelos Ocultos de Markov con DTW (MDTW), Modelos Ocultos de Markov con DTW aproximados por izquierda y derecha (MID) y Modelos Ocultos de Markov con relleno de palabras (MRP)).

El conjunto de algoritmos integrados en la interfaz final posee la flexibilidad suficiente para ampliar la cantidad de palabras de prueba de dicción y, más aún, creemos que se pueden ampliar las pruebas de dicción a otros idiomas modificando el corpus de voces al idioma que se deseé.

7. Bibliografía y Referencias

[1] A. Pedroza, J. de la Rosa, “El invisible y asombroso proceso de la comunicación oral: bases sobre reconocimiento de voz”. Pistas Educativas. No. 112. Noviembre 2015. Pp. 1310-1330.

[2] B. Plínio, "On the Defense of von Kempelen as the Predecessor of Experimental Phonetics and Speech Synthesis Research". The Ninth International Conference on the History of the Language Sciences. 2007. Pp. 101-106.

[3] E. David, O. Selfridge, "Eyes and Ears for Computers". Proceedings of the IEEE. Vol.50. Mayo 1962. Pp. 1093-1101.

[4] B. Gold, N. Morgan, D. Ellis, Speech and audio signal processing: Processing and Perception of Speech and Music. 2da. Edición. 2011. Editorial WILE. 688 páginas.

[5] C. Seelbach, “A perspective on early commercial applications of voice-processing technology for telecommunications and aids for handicapped". Human-Machine Communication by Voice. 1993. Pp. 9989-9990.

[6] C. Caballero, Cómo educar la voz hablada y cantada. 8va. Edición. 1994. Editorial EDAMEX. 257 páginas.

[7] L. Beltrán, Simulación de modelos ocultos de Markov aplicados al reconocimiento de palabras aisladas, utilizando el programa Matlab. Tesis de Licenciatura. Escuela Politécnica Nacional: Escuela de Ingeniería. Quito. 2003.

[8] L. Rabiner, “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. Proceedings of the IEEE. Vol. 77. No. 2. Febrero de 1989. Pp. 257-286.

[9] H. Silva, Reconocimiento Automático de locutor y realización de un sistema experimental. Tesis de Maestría. Centro de Investigación Científica y de Educación Superior de Ensenada. 1994.

[10] S. Prasad, T. Kishore, "Hybrid HMM/DTW based Speech Recognition with Kernel Adaptive Filtering Method”. International Journal on Computational Sciences & Applications (IJCSA). Vol.1. No.4. Febrero de 2014. Pp. 11-21.

[11] Reconocimiento de voz para la aplicación en domótica. Universidad Tecnológica Nacional: Facultad Regional San Nicolás. 2008.

[12] CRIBEO. http://www.cribeo.com/ocio y cultura/1004/"las-10-palabras-mas-del-espanol. Junio de 2014.

[13] K. Murphy, Hidden Markov Model (HMM) Toolbox for Matlab [Algoritmo Computacional]. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html. Mayo de 2015.

8. Autores

M.C. Ángel David Pedroza Ramírez obtuvo el grado de Maestro en Ciencias de la ingeniería por la Universidad Autónoma de Zacatecas en 2015. Actualmente se encuentra cursando el segundo semestre del Doctorado en Ciencias de la Ingeniería con especialidad en Procesamiento de Señales y Mecatrónica en la

misma institución. Su línea de investigación actual es en reconocimiento de voz así como identificación automática enfocada en bioacústica.

Ph.D. José Ismael de la Rosa Vargas obtuvo el grado de Doctor en Ciencias con especialidad en Procesamiento de Señales y Control (noviembre de 2002), por parte de la Universidad Paris Sud (XI) y de la Escuela Superior de Electricidad (SUPELEC) al sur de Paris (Gif-sur-Yvette), Francia. Trabaja actualmente en procesamiento de imágenes y voz, métodos estocásticos en problemas inversos e instrumentación.

M. en C. Ernesto García Domínguez obtuvo el grado de Maestro en Ciencias con especialidad en Electrónica y Telecomunicaciones (área de Instrumentación) en julio de 1993 por parte del Centro de Investigación y de Educación Superior de Ensenada (CICESE), en Baja California.

Ph.D. Hamurabi Gamboa Rosales obtuvo el grado de Doctor en Ciencias con especialidad en Ingeniería Eléctrica en el área de Procesamiento de Señales en 2009, en el Institute of Acoustics and Speech Communication de la Universidad Tecnológica de Dresden, Alemania.

M.C. Aldonso Becerra Sánchez obtuvo el grado de Maestro con orientación en Computación, de la Universidad Autónoma de Zacatecas en 2007. Actualmente se encuentra estudiando el Doctorado en Ciencias de la Ingeniería en la misma Institución, además de ser docente en dicha Facultad.