El invisible y asombroso proceso de la comunicación oral: bases sobre reconocimiento de voz

Ángel David Pedroza Ramírez

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Carretera a la Bufa No.1 Col. Centro Zacatecas, Zac., Teléfono: (492) 92 296 99

P.A.D_16@hotmail.com

José Ismael de la Rosa Vargas

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Av. López Velarde No. 801 Col. Centro Zacatecas, Zac., Teléfono: (492) 925 66 90, ext. 3956

joseismaelrv@gmail.com, ismaelrv@ieee.org

Resumen

La comunicación, cuyo fin primario es la transmisión de información, forma parte fundamental de las necesidades básicas del ser humano. El proceso de la generación del habla y más aún el de la comunicación, es mucho más complejo de lo que se podría llegar a creer dado el nivel de coordinación que se requiere para producir, transmitir y decodificar un mensaje. El reconocimiento de voz se basa en el estudio sobre el proceso del habla y la comunicación, y la forma en que este conocimiento puede ser aplicado.

El presente documento resume y brinda una revisión sobre el estado del arte y las bases para entender el reconocimiento de voz desde el punto de vista fisiológico y como una rama de la ciencia ampliamente utilizada en la tecnología de uso cada vez más común hoy en día.

Palabra(s) Clave(s): comunicación hombre-máquina, comunicación oral, habla, voz.

1. Introducción

La comunicación oral es una de las herramientas de expresión y comunicación más importante que en el pasado dio paso hacia la evolución y que en el mundo actual, forma parte de una necesidad básica.

El proceso de comunicación oral (entre dos personas) está formado por tres elementos principales: Locutor, medio y oyente. El locutor, mediante una serie de procesos de coordinación de órganos, articula las palabras (previamente procesadas por el cerebro) para formar un mensaje. Dicho mensaje se transmite por un medio (generalmente aire) para llegar al oyente (receptor) quien lo decodifica y así cierra el ciclo o cadena del habla.

El reconocimiento de voz es una rama de la investigación multidisciplinaria y que ha logrado dar solución a diversas problemáticas del mundo real que van desde la ayuda a personas con deficiencias hasta los complejos sistemas de navegación con tecnología GPS. 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 (como es el caso de la realización de pruebas de dicción).

En las siguientes secciones se dará una breve reseña sobre los aspectos más importantes que fundamentan al reconocimiento de voz aunado a aspectos relacionados con el estado del arte del área de investigación, pasando por una reseña histórica en la sección 2, luego presentando algunos aspectos formales sobre modelos de producción de voz en la sección 3, hasta llegar a la sección 4 en donde se presentan los aspectos más relevantes sobre el reconocimiento de voz y cómo se implementa para la comunicación hombre–máquina. Finalmente, en la sección 5 se presentan

algunas conclusiones sobre el diseño de sistemas de reconocimiento y algunos comentarios sobre el trabajo que actualmente realizamos.

2. Reseña histórica

Durante millones de años los seres vivos han utilizado la comunicación para diversas necesidades y de dicha forma transmitir información para lograr interactuar con el exterior.

El ser humano ha tenido la necesidad de transmitir ideas, sentimientos y pensamientos dada la necesidad de vivir en sociedad. En un inicio la comunicación se hacía solo por medio de sonidos o de forma pictográfica, sin embargo, no fue hasta que se logró tener una comunicación con sonidos articulados que tomó la importancia que tiene hoy en día.

La información que se transmite a través de la voz posee características que pueden ser analizadas por medio de sistemas de reconocimiento de voz. Dicha rama de la ciencia se subdivide en reconocimiento del habla y reconocimiento del locutor.

Entre las aplicaciones donde se encuentra el reconocimiento de voz en la actualidad se encuentran telefonía, Seguridad, Manos libres, etc.

Al principio los sistemas de reconocimiento de voz se enfocaron en la producción sintética de voz, como fue el caso de la máquina creada por Von Kempelen como lo describe [1], la cual, mediante un arreglo mecánico, producía sonidos parecidos a los producidos por el ser humano, sin embargo este sistema requería de horas de entrenamiento. Este novedoso pero burdo invento daría más tarde lugar a los sintetizadores musicales y a los órganos que utilizaban tarjetas perforadas para producir ciertas melodías.

Poco a poco los sistemas fueron haciéndose más sofisticados hasta que años más tarde Thaddeus Cahill descubrió que cualquier sonido podía ser sintetizado como una

sumatoria de pequeñas ondas sinusoidales el cual revolucionó el concepto sobre la síntesis de música. Diversos inventos fueron creados, como el caso del *telharmonium* el cual utilizaba este principio de adición de ondas.

Sin embargo, el primer invento enfocado al reconocimiento de voz fue un juguete llamado *Radio Rex,* como lo indica [2], manufacturado en 1926. Este funcionaba mediante un mecanismo de activación por comando de voz. Al ser nombrada la palabra "Rex", el mecanismo hacía que la energía contenida en la palabra disparara el dispositivo y liberara al perro. Sin embargo, este dispositivo comercial tenía la desventaja que cualquier palabra con la misma energía que la palabra "Rex", hacía que el sistema funcionara, es decir, liberara al perro.

No fue sino hasta 1950 cuando los laboratorios Bell crearon el primer reconocedor más sofisticado hasta esa fecha; logrando reconocer dígitos de un solo hablante. Sin embargo aún era un poco burdo para las necesidades tecnológicas.

El ámbito del reconocimiento de voz siguió desarrollándose y en 1960 aparecen tres técnicas que fueron la base para el reconocimiento de voz (indicadas en [3]):

- Transformada Rápida de Fourier (FFT): Forma eficiente de la transformada de Fourier, interpretada como un banco de filtros.
- Análisis Cepstral: Utilizado como una aproximación espectral.
- Codificación lineal predictiva (LPC): Modelos de auto-regresión para representar la generación de voz.

Adicionalmente se crearon nuevos métodos de aproximaciones:

1. Alineamiento temporal dinámico (DTW por sus siglas en inglés como lo indica [7]): Esquema de optimización, utilizado como método de normalización (puede ser utilizado para corregir variaciones por diferentes pronunciaciones de la misma palabra con diferente duración).

2. Modelos ocultos de Markov (MOM o HMM por sus siglas en inglés como lo indica [11]): Modela una secuencia observada como la generada por una secuencia desconocida de variables.

Específicamente, los modelos ocultos de Markov fueron desarrollados por primera vez por un grupo de IBM en 1970 los cuales, en primera instancia, se utilizaron para el reconocimiento continuo del habla.

Para 1976 se incorpora información semántica y de sintaxis con el fin de lograr más allá del reconocimiento del habla la comprensión del lenguaje. Para este fin, se incorporaron y desarrollaron investigaciones con redes neuronales, DTW y HMM.

Los Modelos Ocultos de Markov tuvieron gran impacto hasta mediados de 1980, siendo implementados hasta la fecha en algunos sistemas comerciales (Sistema de reconocimiento de voz de la marca Apple llamado Siri como lo indica en [4]).

A finales de 1980 se utilizó información acústico-fonética para desarrollar reglas de clasificación para sonidos del habla. Para ese mismo año se utilizaron redes neuronales para delimitar secciones de frontera (dentro de la señal de voz, ¿qué parte de la señal posee información y cuál no?). Luego de una serie de avances en materia de bases de entrenamiento y subsistemas de extracción de información de la señal de voz, para 1998 algunos sistemas podían reconocer cerca de 60,000 palabras de vocabulario en tiempo real con menos del 10 por ciento de error.

Hoy en día, y luego de un proceso que como se explicó con anterioridad comenzó con invenciones burdas hasta llegar a los complejos sistemas híbridos, el reconocimiento de voz es un área de investigación multidisciplinaria que tiene muchísimas aplicaciones aún por seguir explorando.

3. Teoría fundamental sobre el proceso de producción de voz

3.1. El sonido

El sonido, como se define en [5], es una vibración (o movimiento ondulatorio) que se transmite en el aire (generalmente), hasta alcanzar el oído de quien lo percibe. Dicha vibración tiene cierta amplitud, frecuencia, duración y forma. Las características de la vibración caracterizan al sonido producido.

Siendo así, la frecuencia corresponde al tono del sonido (también llamado altura). Las frecuencias que son altas son denominadas "frecuencias en tono agudo" y las frecuencias bajas se denominan "frecuencias en tono grave". Es interesante mencionar que la cualidad de percepción la proporciona el tono mientras que el dolor auditivo experimentado en la audición lo da la energía y no, por tanto, el tono que se produce.

Al respecto, el sonido se puede clasificar de la siguiente manera:

- Infrasonidos: Frecuencias menores a 20 Hz.
- Percepción humana: Entre 20 y 20,000 Hz.
- Ultrasonidos: Frecuencias mayores a 20,000 Hz.

La amplitud se corresponde directamente a la intensidad del sonido (volumen) clasificado como "alto" o "bajo". Esta característica se refiere a la distancia entre el extremo superior e inferior de la onda de sonido. Dicho parámetro está relacionado con la intensidad y por tanto con la energía.

Características de la **vibración.**		Características del **sonido**
Frecuencia	produce	Tono
Amplitud	"	Intensidad
Duración	"	Duración
Forma	"	Timbre

Tabla 1. Correspondencia vibración a sonido.

La duración de la vibración es también la duración del sonido. Por otro lado, las vibraciones parásitas permiten identificar a la fuente de sonido (a cada una corresponden diversos armónicos) dado que son producidas por diversas formas de vibración. Esa cualidad es llamada en la teoría del sonido como timbre. Así, el sonido de una flauta y un violín con la misma intensidad, tono y duración se diferencian por el timbre (ver Tabla 1).

Las vibraciones que puede captar el ser humano como el sonido están limitadas por la frecuencia. Inferiores a 30 ciclos por segundo se percibe como movimiento mientras que, si sobrepasan los 18,000 ciclos por segundo rebasa las posibilidades de audición.

Si la amplitud es muy leve no se produce sensación auditiva pero si es muy grande, puede dañar de forma permanente el órgano de la audición. En cuanto a la duración, si es menor a 100/10000 de segundo solo se percibe como un "click" no identificable. Si la forma de vibración y frecuencia permanecen regulares se percibe como sonido; si alguno de los dos es irregular, el resultado es un ruido.

3.2. ¿Cómo se produce la voz?

La voz humana es conocida como el flujo de ondas (de tipo sonoras o silencios) que se propaga por medio del aire (presión de moléculas). El proceso de producción de voz conlleva la combinación de órganos, huesos, músculos y algunos sistemas funcionales (como lo explica [18]). La combinación de estos elementos en conjunto con la postura, la respiración y el estado emocional influyen en dicho proceso de producción (como se puede ver en [6]). Este proceso se resume de la siguiente forma: Previa inhalación de aire en los pulmones, la voz se produce cuando el aire es exhalado; al expandirse el diafragma, los órganos de respiración proporcionan un flujo de aire a los órganos de la fonación, y es ahí donde el sonido adquiere sus características primarias en donde las cuerdas vocales juegan un rol importante. Luego, se agregan otras características impuestas por los órganos articulatorios, brindando al final una señal acústico-fonética.

El ser humano posee un sistema productor de sonido con ciertas características anatómicas (ver Fig. 1). En el sistema fonador interviene no solo la laringe, sino también los pulmones (para proporcionar el aire) y los modificadores de sonido (resonadores faringo-buco-nasales). Es por ello que la calidad de sonidos producidos está directamente determinada por la posición de la lengua, los labios y la mandíbula.

La fonación funciona mediante dos sistemas: Sistemas directos y sistemas indirectos. Los sistemas directos están compuestos por el sistema respiratorio, órganos articulatorios, órganos de resonancia y el aparato fonador. Por otro lado, los sistemas indirectos están conformados por el sistema muscular, el aparato auditivo y el sistema óseo.

La fuente de aire en el sistema de producción de voz se da gracias a los pulmones en la espiración. Según la cantidad de aire que se expulse por los pulmones es la intensidad que se tiene en la voz; mientras que gracias al diafragma se hace posible el variar su duración.

Específicamente el aparato fonador es el encargado de reproducir el sonido y está conformado por la laringe y las cuerdas vocales.

La laringe, además de proteger las vías respiratorias, es una caja sonora encargada de la fonación (donde se origina el tono fundamental de la voz). Entre las funciones más importantes de la laringe, como lo indica [6], se encuentran la respiratoria (permite entrada y salida del aire al respirar), esfinteriana (permite realizar esfuerzos), deglutatoria (impide el paso de comida a los pulmones) y de fonación (produce la voz gracias a las cuerdas vocales).

Las cuerdas vocales se encuentran dentro de la laringe y son membranas de tejido de la laringe. Estas producen el sonido mediante la vibración al paso del aire y la aproximación o separación entre si producen voz o silencio respectivamente. Su forma, aunado a las de la garganta, nariz y boca, es decir, las cavidades resonantes, determinan el sonido de la voz en cada persona.

El tono de voz se relaciona con la frecuencia a la que vibran las cuerdas vocales. Específicamente, en los adultos masculinos se presenta una frecuencia baja (120 ciclos por segundo) por lo que poseen una voz más grabe. En el caso de las mujeres (250 ciclos por segundo) se presenta una voz aguda; en los niños y jóvenes es aún más aguda (400 ciclos por segundo). Es necesario aclarar que, aunado a esto, la longitud, tensión y masa de las cuerdas vocales influyen en el tono de voz. Por ejemplo, las cuerdas vocales más gruesas vibran más despacio por lo que producen un sonido más grave.

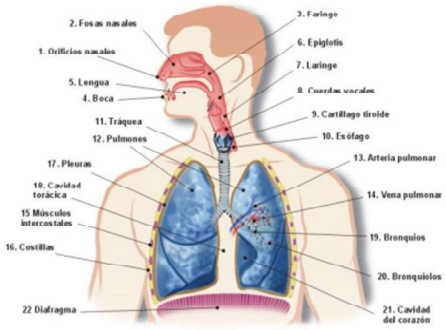

Fig. 1. Sistema Fonador.

En relación al aspecto psicológico y la voz, este parámetro es notorio en la respiración. La respiración varía en un ritmo particular en proporción al estado de ánimo (acelerado ante la excitación y limitado ante la tristeza).

3.3. Clasificación de los fonemas

En cuanto a la forma en que se clasifica cada sonido, cada lengua o lenguaje posee un conjunto propio de sonidos que el tracto vocal permite producir con naturalidad. A las unidades sonoras propias de un idioma se les llama "fonemas" (representados entre diagonales) y varían según las características impuestas por cada idioma. La combinación entre fonemas genera sílabas y a su vez, un conjunto de sílabas forma una palabra.

Una de las características más importantes de los fonemas es que, a un enunciado hablado, se le puede descomponer en estas unidades sonoras y si se cambia o sustituye por otro se cambia el contenido lingüístico. Se debe distinguir entre fonemas y letras ya que los fonemas son elementos sonoros de un idioma; mientras que las letras son las representaciones gráficas de los fonemas. En algunas ocasiones un mismo fonema es representado de diversas formas como en el caso de la representación de k, c y q para el mismo sonido.

Las frecuencias formantes permiten clasificar los diferentes sonidos o fonemas. Dichas formantes dependen de la dimensión y forma interna de la boca o tracto vocal (cada configuración provee unas formantes particulares como lo indica [7]).

En cuestiones de modelado, solo es necesario tomar en cuenta las formantes F1, F2 y F3. Las dos primeras formantes las determina la posición de la lengua; cuanto más baja esté, F1 posee una frecuencia más alta. Para el caso de F2, su frecuencia es mayor cuanto más hacia adelante esté la lengua.

Por otro lado, la frecuencia fundamental (pitch) o F0 corresponde a la frecuencia de oscilación de las cuerdas vocales en los sonidos sonoros. Si la frecuencia fundamental es mayor que la de la frecuencia de las formantes se hace difícil de distinguir y por ende difícil al reconocer.

Para el caso del idioma español la subdivisión de los fonemas es la siguiente:

1. Timbres Básicos,
2. Sonidos auxiliares.
3. Ataques:
    - Fuertes,
    - Suaves,
    - Sonidos Mixtos.

Timbres Básicos: Producidas con la intervención de las cuerdas vocales. Se denominan como timbres básicos debido a que son variaciones del sonido producido naturalmente por las cuerdas vocales y la columna de aire que se produce mediante la respiración. En esta clasificación se encuentran las vocales.

Para el caso del idioma español las vocales son: /a/, /e/, /i/, /o/ y /u/. De entre ellas la que se produce con naturalidad es la /a/. En cualquier otra, la posición de la lengua se modifica y por ello se crea otra clasificación como lo indica [8]:

i. Vocales Palatales: La lengua se expone gradualmente y se eleva hacia el paladar. En esta clasificación se encuentran las vocales /e/ e /i/.
ii. Vocales velares: La lengua se contrae y se eleva al velo del paladar. En esta clasificación se encuentran las vocales /o/ y /u/.

De modo general, las vocales se pueden clasificar de acuerdo al grado de abertura de la boca, posición de la lengua y grado de sonoridad (ver Fig. 2). La correcta producción de ellos (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. La vibración no se produce en la laringe sino en el paladar blando, la lengua, etc. Corresponden a las consonantes sonoras: /m/, /n/, /l/, /s/, /j/, /r/.

Ataques: No son propiamente un sonido dado que no posee una vibración periódica. Son modos de iniciar un timbre básico o un sonido auxiliar y representan las distintas formas de la oclusión que impide fluir al aire. Una vez liberada, la columna de aire vibra produciendo un timbre básico o un sonido auxiliar; en ese momento el ataque deja de existir. Su naturaleza es momentánea y requiere de su identificación perfecta.

Fig. 2. Triángulo vocálico del español (HELWAG).

Pueden ser suaves o fuertes según la brusquedad de la liberación de aire. Ellos son: /b/, /p/, /d/, /t/, /k/. Los sonidos mixtos, como caso particular de ataques, se refiere a la sucesión de sonidos auxiliares /l/ o /n/ y el timbre básico /i/, los cuales producen los fonemas /ll/ y /ñ/; o las combinaciones del ataque /k/ con el sonido auxiliar /s/ produciendo /x/.

La calidad de la dicción se fundamenta en la correcta articulación de los ataques. Por otro lado, algo que debe ser tomado en cuenta para poder hablar, es respetar las reglas que posee cada idioma para formar combinaciones de fonemas. Por ejemplo, para el caso del castellano, los sonidos básicos y los sonidos auxiliares son combinables con cualquier fonema. Los ataques casi nunca son combinables entre si y no pueden producirse aislados; requieren de la combinación con un timbre básico o un sonido auxiliar.

4. ¿Qué es el reconocimiento de voz?

El reconocimiento de voz se basa en el conocimiento sobre el proceso del habla y la estructura que conlleva el lenguaje en el ser humano. Desde esta perspectiva se tiene que el fin último del reconocimiento de voz es hacer que las computadoras (máquinas) logren un nivel suficiente para expresar y comprender la comunicación como lo hace en su naturaleza el ser humano. Actualmente las empresas más importantes en desarrollo de aplicaciones en sistemas de reconocimiento de voz, como en [9], son Philips,

Lernout & Hauspie, Sensory Circuits, Dragon Systems, Speechworks, Vocalis, Dialogic, Novell, Microsoft, NEC, Siemens, Intel.

El reconocimiento automático de voz (RAV) se divide en dos áreas, como se indica en [7], enfocadas cada una en una tarea específica:

I. Reconocimiento automático del habla (RAH).
II. Reconocimiento automático de locutor (RAL), el cual se subdivide a su vez en:
    1. Identificación automático de locutor (IAL).
    2. Verificación automática del locutor (VAL).

Los sistemas de reconocimiento se pueden además clasificar como dependientes e independientes del locutor. Para el caso de los sistemas dependientes del locutor (cooperativo), el mensaje que se identifica o articula es una palabra o mensaje fijo (frase pre-establecida); mientras que para el caso en el que son independientes del locutor (no cooperativo), se posee la libertad para articular cualquier mensaje para ser reconocido.

Debe aclararse que para este último caso, la librería es mucho más extensa en relación al vocabulario que el locutor posee por lo que se puede decir que tiene la "libertad" para pronunciar cualquier mensaje. La eficiencia de los sistemas dependientes es mayor que la de los sistemas independientes.

El estudio del reconocimiento de voz se basa en tres principios:

1. La información de la señal de voz se puede representar por el espectro en amplitud a corto plazo de la forma de onda de la voz.
2. El contenido de la voz se puede expresar en forma escrita (una secuencia de símbolos fonéticos o caracteres del alfabeto).
3. Es un proceso cognoscitivo, es decir, la comprensión está ligada a la gramática, semántica y estructura del lenguaje.

La dificultad que se presenta para poder llevar a cabo reconocimiento de voz (en cualquiera de sus áreas) es la variabilidad presente en las señales de voz, es decir, las variaciones que se presentan para cada locutor, las condiciones acústicas, entre otras. Sin embargo, dentro de estas variaciones se debe determinar cuáles de ellas son relevantes y cuáles no, es decir, para cada aplicación determinar los parámetros a tomar en cuenta y los que se pueden despreciar.

En general, entre los problemas y dificultades que se encuentran en las tareas de reconocimiento de voz se consideran:

- El tamaño del vocabulario,
- ambigüedad acústica y grado de confusión,
- calidad ambiental.

Entre las más importantes (existen algunas adicionales).

Existen cuatro enfoques principales, como lo indica [10], que se pueden utilizar para realizar el reconocimiento del habla:

1. Contraste de patrones
    - Supone al habla como una secuencia de palabras cada una con un patrón o conjunto de patrones.
    - Se compara la unidad de habla entrante con los patrones de referencia almacenados.
    - Utilizado en reconocimiento de palabras aisladas o conectadas.
    - Esta técnica no permite generalizar los patrones por lo que es necesario crear una biblioteca para cada hablante lo cual reduce su nivel de aplicación en tareas de reconocimiento avanzadas.

2. Sistemas basados en el conocimiento:
    - Emula y aplica los conocimientos sobre el habla en tareas de reconocimiento que utiliza el ser humano.
    - Usa técnicas con reglas y sistemas expertos, desde el nivel acústico-fonético hasta niveles más complejos.
3. Modelos estocásticos:
    - Hace uso de modelos estocásticos (Modelos Ocultos de Markov) en lugar de modelos determinísticos.
    - Utilizado comúnmente para el reconocimiento de palabras continuas.

4. Modelos neuronales o conexionistas:
    - No posee tantas restricciones como los modelos estocásticos.
    - Sus tiempos de entrenamiento son razonables.
    - Se han utilizado también como parte de fusiones entre varios enfoques.

Es necesario aclarar que el uso de cada uno o combinación entre ellos, dependerá de cuál es el objetivo que el reconocimiento pretende alcanzar.

Para llevar a cabo el reconocimiento de voz es necesario obtener los parámetros que representen la información espectral contenida en la señal de voz. Para ello, es necesario aplicar un procesamiento matemático con el fin de adecuar la señal para poder analizarla posteriormente.

En principio, es necesario aplicar un pre-procesamiento para después extraer los parámetros propios de la señal. Algunos de los principales métodos de análisis que se aplican en el reconocimiento de voz son: Bancos de Filtros, LPC, Cepstrum, etc.

Por otro lado, existen diversas técnicas para llevar a cabo el reconocimiento de voz. Algunas de ellas, están enfocadas al reconocimiento de palabras aisladas y otras más son comúnmente utilizadas en reconocimiento de habla continua. Sin embargo es necesario tomar en cuenta que la elección de ellas dependerá del objetivo final que se

pretende alcanzar. De los enfoques anteriores, técnicas como Cuantificación Vectorial (CV) (como lo indica [12]), Alineamiento Dinámico en el tiempo (ADT o DTW), Modelos Ocultos Markov (MOM o HMM) (como se indica en [13]) y redes neuronales son algunas de las principales utilizadas en tareas de reconocimiento de Voz.

Cada una tiene algunas ventajas y desventajas propias del enfoque que posee cada uno. Una de las técnicas que actualmente se utiliza para llevar a cabo el reconocimiento de voz con un alto grado de eficiencia (por encima del 90%) son los Modelos Ocultos de Markov (posibilitando la creación de sistemas más complejos de reconocimiento como el descrito en [22]).

La ventaja de utilizar esta técnica respecto de otras es que el reconocimiento, al contrario de los métodos de comparación no estadísticos, es que el reconocimiento se hace respecto de un modelo que caracteriza de forma específica a cada palabra. Estos modelos pueden ser creados con características específicas (modelos en base a fonemas, con una dicción específica, en un idioma concreto, etc.) lo que le da la flexibilidad necesaria para ser aplicado en varias áreas. Ejemplo de ello es la investigación realizada sobre, una vez realizado el reconocimiento, cómo adecuar los Modelos de Markov para identificar la dicción presente en muestras de voz y en base a ello poder realizar ejercicios de dicción de diversas palabras en un idioma específico.

La idea básica de los Modelos Ocultos de Markov aplicados en reconocimiento de voz consiste en: i) crear un modelo para cada una de las palabras que se desean reconocer (proceso de entrenamiento) y ii) almacenarlas con sus respectivas etiquetas. Una vez finalizado este proceso, cuando se desea reconocer una nueva palabra, se calcula la probabilidad de que la repetición a reconocer haya sido generada por un modelo específico (proceso de prueba) y en base a esto se identifica la palabra (como lo explica a detalle [19]).

Actualmente se trabajan en conjunto con otras técnicas como Redes Neuronales (base del asistente comercial Siri), Alineamiento Dinámico en el Tiempo (como en [14]), etc.; con la finalidad de mejorar la eficiencia de reconocimiento.

Por otro lado, las redes neuronales (Neural Networks - NN) consisten en unidades de cálculo simple interconectadas (como lo indica [17]). Estas pretenden interconectar un conjunto de unidades de proceso (o neuronas) en paralelo de forma similar a como lo hace el ser humano (como lo indica [16]) (obteniendo también prestaciones similares en reconocimiento tanto en tiempo de respuesta como en tasa de error). Este método es útil cuando se desean evaluar varias hipótesis en paralelo (como lo indica [15] y [23]) o como herramienta de ayuda en algunas aplicaciones (como la descrita en [21]).

La idea básica detrás de una red neuronal es: dados una serie de parámetros, combinarlos con la finalidad de predecir un cierto resultado. Las unidades de proceso son de varios tipos pero la más utilizada dispone de varias entradas y la salida es el resultado de una transformación no lineal a la combinación lineal de las entradas.

En cuanto al tipo de red, son definidas en cuanto a la forma en que se conectan las neuronas, el tipo de neurona que lo conforma y la forma de entrenamiento de la red (como se explica en [20]).

5. Conclusiones

De la presente investigación se pueden obtener las siguientes conclusiones:

- El reconocimiento de voz es un área de investigación con bases bien fundamentadas y que ha dado la pauta para la solución de problemas en el mundo actual y para la comunicación hombre-máquina.

- Aún con las dificultades que se presentan para lograr un reconocimiento óptimo, es necesario saber notar cuales de éstas características son relevantes y cuáles no, dependiendo de la aplicación a la que se enfoque el reconocimiento de voz.
- Las escuelas de reconocimiento del habla existentes hasta hoy en día han respondido a una necesidad tecnológica presente en cada una de sus épocas de desarrollo, sin embargo, la elección de cada uno de los métodos que se debe utilizar responde al tipo de aplicación que se desea realizar.
- De entre las técnicas para reconocimiento de voz, los Modelos Ocultos de Markov conforman una buena herramienta capaz de ser aplicada en diversas áreas gracias a la flexibilidad con que se pueden diseñar, particularmente en tareas de mejora de dicción que es uno de nuestros objetivos.
- Independientemente de la técnica que se utilice, los métodos de reconocimiento siempre tendrán errores en mayor o menor proporción debido a que se basan en aproximaciones de la señal a identificar.

6. Referencias

[1] 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. 101-106 pp.

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

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

[4] R. Esparza, " Cómo funciona Siri". Como funciona: Edición México. No. 4. 2014. 45 p.

[5] C. Cristían, La voz hablada y cantada. 8va. Edición. 1994. Editorial EDAMEX. 257 pp.

[6] G. de las Heras, L. Rodríguez, Materiales para cuidar mi voz. Fundación MAPFRE-UCLM. 44 pp.

[7] 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.

[8] 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.

[9] J. Flores, Técnicas para el reconocimiento de voz en palabras aisladas en la lengua náhuatl, Tesis de Maestría. Centro de Investigación en Computación. México. D.F. 2009.

[10] A. Ramírez, Reconocimiento automático del locutor mediante técnicas dependientes e independientes del vocabulario para un sistema acotado por el ancho de banda telefónico y realización de un sistema experimental. Tesis de Maestría. Centro de Investigación Científica y de Educación Superior de Ensenada. 1996.

[11] L. Rabiner, B. Juang, Fundamentals of speech recognition. 1993. Editorial Prentice Hall. 507 pp.

[12] A. Buzó, A. Gray, R. Gray, J. Markel, “Speech Coding Based Upon Vector Quantization”. IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol. assp-28. No. 5. Octubre de 1980. 562-574 pp.

[13] 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. 257-286 pp.

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

[15] J. Varela, J. Loaiza, Reconocimiento de palabras aisladas mediante redes neuronales sobre FPGA. Tesis de licenciatura. Facultad de Ingenierías: Eléctrica, Electrónica, Física y de Sistemas. Universidad Tecnológica de Pereira. 2008.

[16] S. Rascón, Reconocimiento de voz para un control de acceso mediante una red neuronal de retropropagación. Tesis de licenciatura. Escuela superior de ingeniería mecánica y eléctrica Unidad Culhuacan. México. D.F. 2009.

[17] J. Pech, Desarrollo de un sistema de reconocimiento de voz para el control de dispositivos utilizando mixturas gaussianas. Tesis de Maestría. Instituto Politécnico Nacional. Centro de Investigación en Computación. México. D.F. 2006.

[18] J. Rodríguez, Sistema de reconocimiento del locutor basado en modelado no paramétrico. Tesis de Maestría. Instituto Politécnico Nacional. Escuela Superior de Ingeniería Mecánica y Eléctrica. México. D.F. 2008.

[19] G. Pérez, Herramientas de Segmentación y Evaluación de Series Temporales Basadas en Modelos Ocultos de Markov. Tesis de Licenciatura. Universidad Carlos III de Madrid. España. Madrid. 2010.

[20] J. Schmidhuber, "Deep learning in neural networks: An overview". Neural Networks. No. 61. Enero de 2015. 85-117 pp.

[21] A. Abad et al., "Automatic word naming recognition for an on-line aphasia treatment system". Computer Speech and Language. No. 27. Septiembre de 2013. 1235-1248 pp.

[22] A. Hussen, S. Zeiler, D. Kolossa, "Learning Dynamic Stream Weights For Coupled-HMM-Based Audio-Visual Speech Recognitio ". IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 23. No. 5. Mayo 2015. 863-876 pp.

[23] P. Cardinal, P. Dumouchel, G. Boulianne, "Large Vocabulary Speech Recognition on Parallel Architectures". IEE Trasactions on audio, speech, and languaje processing. Vol.21. No. 11. Noviembre de 2013. 2290-2300 pp.

7. Autores

Ph.D. José Ismael de la Rosa Vargas es ingeniero en Comunicaciones y Electrónica egresado de la Universidad Autónoma de Zacatecas en el año de 1995. Obtuvo el grado de Maestro en Ciencias con especialidad en Sistemas Digitales en el área de

Procesamiento Digital de Señales (PDS) en mayo de 1998 por parte del Centro de Investigación y Desarrollo de Tecnología Digital (CITEDI) del Instituto Politécnico Nacional situado en Tijuana, Baja California. Posteriormente obtiene 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.

Ing. Ángel David Pedroza Ramírez es Ingeniero en Comunicaciones y Electrónica por la Universidad Autónoma de Zacatecas. Actualmente se encuentra cursando el último semestre de la Maestría 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 enfocada en pruebas de dicción mediante Modelos Ocultos de Markov.

M.C. Ernesto García Domínguez es ingeniero en Comunicaciones y Electrónica egresado de la Universidad Autónoma de Zacatecas en el año de 1989. 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.