VOZ, UNA NUEVA INTERFAZ HOMBRE-MAQUINA

JORGE AUGUSTO JARAMILLO M. *

RESUMEN

El siguiente escrito ha sido elaborado con base en los documentos actuales y en los esfuerzos hechos por diferentes personas y entidades gubernamentales, en implementar sistemas que permitan incorporar el habla humana en un computador, como una interfase alternativa y más natural.

Aquí se da un esbozo bastante global de lo que es esta interfase, pues los estudios hechos al respecto van desde simples algoritmos para el análisis de las señales (determinación de las características implícitas en ellas) hasta sistemas bastante complejos que desean establecer una comunicación continua con cualquier locutor.

La primera parte del escrito trata el funcionamiento de los aparatos transductores humanos. Una segunda parte da a conocer los diferentes enfoques al reconocimiento del habla. Una tercera presenta algunas de las posibles aplicaciones del reconocimiento de Voz. En seguida se analiza la forma cómo se genera la Voz desde el computador y por último se plantea el modelo general de la interfase Hombre-Máquina.

INTRODUCCION

Nos encontramos en plena transición hacia un mundo en el cual la comunicación entre el hombre y la máquina es más natural. Los sistemas de computadores que entiendan el lenguaje de los usuarios, modificarán en forma radical, tanto las interfases "Hombre-Máquina" como las comunicaciones entre las personas. El Reconocimiento Automático del Habla por computador es un campo multidisciplinario, especialmente relacionado con la inteligencia artificial y el procesamiento digital de señales. Su objetivo es la concepción y realización de sistemas automáticos capaces de interpretar la señal vocal procedente de algún locutor humano. De la misma forma como el computador podría interpretar estas señales del habla, también es necesario que este genere su propia Voz para poder establecer una comunicación más natural e intuitiva.

1. APARATOS TRANSDUCTORES HUMANOS

Para el entendimiento de los procesos que implican el reconocimiento y la generación de voz por máquina, es necesario revisar los conceptos y entender el funcionamiento de los aparatos transductores humanos que realizan estas tareas.

El habla es uno de los más útiles y complejos medios de comunicación en el hombre. La señal sobre la que descansa este medio (VOZ) se transmite principalmente a través de un canal que está constituido por ondas de presión que se propagan a través del aire. Los aparatos transductores humanos que producen y captan la señal son:

1.1. APARATO FONADOR

El Aparato Fonador está formado básicamente por tres elementos (Fig 1.): 1) Un Generador de energía *(pulmones)*, 2) Un Sistema vibrante *(Laringe y Cuerdas Vocales)*, y 3) Una cavidad resonante *(Conducto Vocal)*.

El sonido se genera mediante la transformación de una corriente de aire que es expulsada desde los pulmones proporcionando una diferencia de presión necesaria para crear un flujo de aire que activará a la laringe y las demás cavidades del tracto vocal. Para esta transformación intervienen órganos que realizan

* Ingeniero de Sistemas Univ. Autónoma de Colombia, Auxiliar de Docencia Facultad de Ingeniería de Sistemas E.A.N.

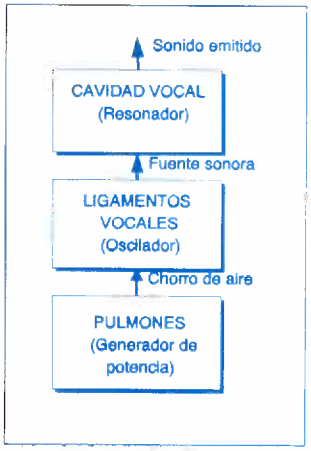

Fig. 1 Diagrama del funcionamiento del Aparato Fonador.

otras tareas como la respiración, ingestión y digestión de los alimentos.

1.2. APARATO AUDITIVO

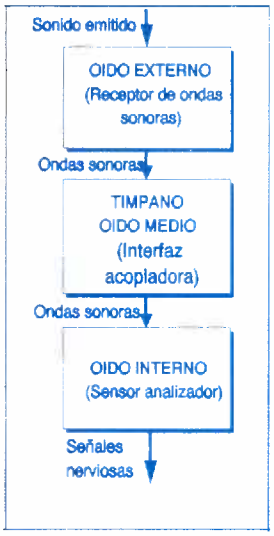

Fig. 2 Diagrama del funcionamiento del Aparato Auditivo.

Está formado por tres elementos básicos (Fig. 2.): 1) Un pabellón colector de las ondas sonoras *(pabellón auditivo y conducto externo)*, Una interfaz acopladora de impedancias acústicas *(tímpano y oído medio)* 3) Un sistema sensor-analizador frecuencial *(caracol y membrana basilar)*.

Desde el punto de vista de la comunicación oral los dos primeros son de poco interés, por considerarse como simples prolongaciones del canal de transmisión del aire. El tercero, presenta gran interés por desarrollarse en él procesos pocos conocidos de análisis de la señal acústica.

2.ENFOQUES AL RECONOCIMIENTO AUTOMATICO DEL HABLA

Se entiende como enfoques, a los diferentes modelos planteados para dar tratamiento al complejo análisis de las señales del habla.

2.1.Enfoque Acústico-Fonético

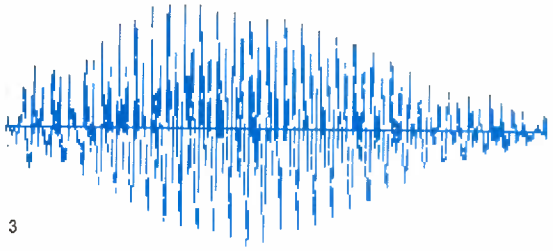

Fig. 3

El enfoque Acústico-Fonético, es sin duda un método viable y ha sido ampliamente estudiado por más de 40 años, apuntando hacia un modo directo [2] de reconocimiento del habla por una máquina (computador).

El enfoque acústico-fonético está basado en la teoría de la fonética acústica y postula que existen unidades distintivas fonéticas finitas en el lenguaje hablado y que las unidades fonéticas están caracterizadas ampliamente por un conjunto de propiedades[3] que pueden manifestarse en la señal del habla, o en su espectro de onda sobre el tiempo. Aún, cuando las propiedades acústicas de las unidades fonéticas son altamente variables, se asume que las reglas que gobiernan la variabilidad, son directas y pueden ser rápidamente aprendidas y aplicadas en situaciones prácticas.

Sin embargo, por una variedad de diferentes razones, el enfoque acústico-fonético no ha logrado el mismo éxito en sistemas prácticos que tienen métodos alternativos. Algunas de estas razones son:

1. El método requiere un amplio conocimiento de las propiedades acústicas de las unidades fonéticas. Este conocimiento, es a lo mejor incompleto, y en el peor de los casos no disponible para las situaciones más simples.
2. La escogencia de las características es realizada en su mayor parte por consideraciones ad hoc. Para sistemas de mayor envergadura la escogencia de las características está basada en la intuición y no es óptima en casos variables o bien definidos.
3. El diseño de un clasificador de sonidos tampoco es óptimo. Métodos ad hoc son generalmente usados para construir árboles binarios de decisión.
4. Los procedimientos automáticos existentes, no son bien definidos acorde con los sistemas reales, en la rotulación[4] del habla. En realidad, no hay un camino ideal para la rotulación y segmentación en el entrenamiento del habla de una manera uniforme y consistente.

2.2. Reconocimiento de patrones

El enfoque del reconocimiento de patrones al reconocimiento del habla es básicamente uno en el cual los patrones del habla son utilizados directamente sin determinación de las características explícitas de la señal (en el sentido acústico-fonético) y segmentación[5]. El método tiene dos pasos: la enseñanza y el reconocimiento de patrones. El "conocimiento" del lenguaje es aprendido por el sistema a través del

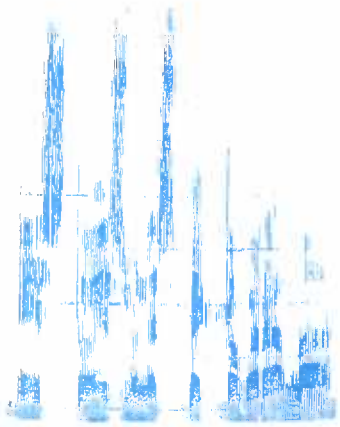

Fig. 4

procedimiento entrenador. El concepto es, que si versiones suficientes de un patrón a ser reconocido (sea un sonido, una palabra, una frase, etc.) son incluidos en un conjunto entrenador provisto al algoritmo, el procedimiento entrenador [6] debe poder caracterizar adecuadamente las propiedades acústicas del patrón (sin consideración para el conocimiento de cualquier otro patrón presentado al procedimiento entrenador).

Este tipo de caracterización del habla a través del entrenamiento es llamado clasificación de patrones debido a que el computador aprende cuales propiedades acústicas de la palabra hablada, son confiables y repetibles a través de todos los símbolos entrenadores del patrón. La utilidad del método es la etapa de comparación de patrones, que hace una comparación directa de la señal (la palabra hablada a ser reconocida), con cada patrón posiblemente aprendido en la fase entrenadora y clasifica la palabra hablada desconocida de acuerdo con los patrones similares. El enfoque del reconocimiento de patrones es el método opcional para el reconocimiento de palabras habladas por tres razones:

1. La simplicidad de uso. El método es fácil de entender, es rico en justificación teorica de la comunicación y matemática, para procedimientos individuales utilizados en entrenamiento y decodificación, y está siendo entendido y utilizado ampliamente.
2. Robustés e invariancia ante diferentes vocabularios del lenguaje, usuarios, grupos de características, algoritmos de comparación de patrónes y reglas de decisión. Estas propiedades lo hacen el algoritmo apropiado para un rango amplio de unidades del habla (variando desde unidades de fonemas a todo el camino a través palabras, frases, y las oraciones), vocabularios de palabras, poblaciones de oradores, condiciones de transmisión, etc .
3. Alto rendimiento garantizado. El enfoque del reconocimiento de patrones al reconocimiento del habla provee consistentemente alto rendimiento en cualquier tarea, que es razonable por la tecnología y provee un camino claro para extender la tecnología a un amplio rango de direcciones, tal que el desempeño disminuye tanto como el problema se ponga más y más difícil.

Debido a que el sistema es no sensible a las clases de sonidos, las técnicas básicas de patrones, son aplicadas a un amplio rango de sonidos del habla, incluyendo frases, palabras aisladas y unidades fonéticas. De aquí veremos como un conjunto de estas técnicas desarrolladas para una clase de sonido (por ejemplo, palabras) pueden generalmente ser aplicadas directamente a diferentes clases de sonidos (por ejemplo, a unidades fonéticas) con muy pequeña o ninguna modificación a los algoritmos.

En términos generales el enfoque del reconocimiento de patrones da la pauta para aplicaciones en sistemas prácticos, debido a su rapidez en el reconocimiento y es relativamente fácil su implementación y mantenimiento.

2.3. Inteligencia Artificial

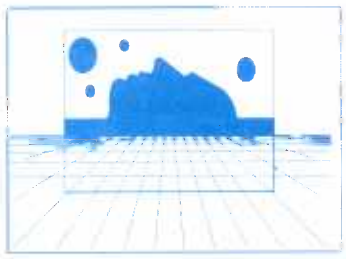

Fig. 5

El así llamado enfoque de la inteligencia artificial al reconocimiento del habla es un híbrido entre el enfoque acústico-fonético y el enfoque de reconocimiento de patrones en el que explotan ideas y conceptos de ambos métodos.

El enfoque de la inteligencia artificial intenta mecanizar el procedimiento del reconocimiento de acuerdo con el modo en que una persona aplica su inteligencia al visualizar, analizar, y hacer finalmente una decisión en los caracteres distintivos acústicos medidos. En particular, entre las técnicas utilizadas dentro de esta clase de métodos, son el uso de un sistema experto para la segmentación y rotulación[7] de modo que este paso crucial y más difícil puede ser desempeñado con algo más que simplemente la información acústica utilizada por métodos acústicos-fonéticos puros (métodos que integren fonemas, léxico, sintaxis, semántica, y uniforme conocimiento pragmático en el sistema experto que ha sido propuesto y estudiado); aprendidos y adaptados sobre el tiempo (por ejemplo, el concepto de que ese conocimiento está frecuentemente adaptado tanto a los modelos estáticos como dinámicos y se tienen que acoplar al componente dinámico de la información).

3. APLICACIONES DEL RECONOCIMIENTO DE VOZ

Aunque los primeros trabajos sobre la parametrización de la voz se hicieron (y continúan haciéndose) con el propósito de reducción de la anchura de banda de los canales de transmisión, pronto surgieron aplicaciones potenciales de sistemas de reconocimiento y síntesis del habla, todas ellas relacionadas más o menos directamente con la comunicación hombre-máquina.

Primer grupo (Aplicaciones clásicas): tareas en las que el operador tiene ocupadas las manos y la vista, pero en las que la interacción con la máquina puede hacerse cómodamente mediante la voz.

Control de calidad.
Manejo y clasificación automática de paquetes, piezas, etc.
Control de máquinas-herramientas.

Segundo grupo de aplicaciones: Aquellas en las que el sistema oral tiende a sustituir totalmente al operador humano en tareas rutinarias.

Respuesta e información telefónica automática.
Servicios de documentación a partir de un banco de datos.
Distribución de llamadas telefónicas en una central.

Tercer grupo: De gran interés social que incluye sistemas de ayuda a los minusválidos.

Control por voz de sillas de ruedas.
Sistemas de aprendizaje del habla para sordomudos.
Prótesis de palabra sintética para ayuda a los individuos con trastornos del habla.

Cuarto grupo: los sistemas de comunicación oral con los computadores en un futuro próximo, pasarán a formar parte de los equipos periféricos estándar.

Pantallas alfanuméricas (y/o) con voz sintética.
Dispositivos de entrada oral de datos.
Lenguajes de mandatos orales.
Traducción simultánea por computador.

4. SINTETIZADORES DE VOZ

La síntesis del habla se utiliza para generar señales del habla en los sistemas de salida vocal. Hay dos técnicas comúnmente utilizadas para este propósito: La primera está basada en tablas y trabajan como un diccionario. Cada palabra está almacenada en tanto una versión de texto como una versión de "sonido".

Una búsqueda es ejecutada en la versión de texto y la hace corresponder con la versión de "sonido". La segunda técnica está basada en reglas, es decir que no es necesario el almacenamiento y las grabaciones de la voz, justamente las reglas son utilizadas para transformar el texto en habla. La utilidad de estas reglas es el de convertir el texto a un conjunto de "descriptores de sonidos" los cuales son llevados a su vez a patrones digitales que describen la señal de salida que oímos.

4.1. MODELO LPC - Para síntesis de VOZ (Linear PredictivO Coding)

Los algoritmos más utilizados para sintetizar la voz, modelan por medio de filtros, sonidos provenientes de una fuente adecuada, de forma tal que solo se requieren muy pocos parámetros para generar la señal. Un ejemplo típico de esta clase de codificación vocal es la técnica de predicción lineal (LPC). El sistema LPC intenta imitar la generación de la voz humana articulando formantes sonoros y áfonos para luego aplicar un filtro variable en el tiempo que corresponde a la estructura resonante de la cavidad vocal y nasal.

En la Fig.6. se muestra el proceso de generación de Voz por el modelo LPC.

La sonoridad se logra generando un tren de impulsos; la afonía, con ruido blanco. Controlando el volumen y los coeficientes de los filtros, el espectro blanco se transforma en otro de señales vocales. El flujo de los parámetros requeridos se obtiene de la secuencia de textos ingresados, aplicando un diccionario que contiene parámetros de los morfemas[8] y una complicada red de reglas de filtrado para obtener la duración y la entonación necesarias. Además, el texto de la frase controla la variación de la altura tonal. Si bien el lenguaje generado de esta forma resulta inteligible, aún no tiene el sonido de una voz natural. Es decir, todavía se requieren reglas de conformación y excepción más precisas y los morfemas del diccionario tienen que moderarse con la mayor exactitud.

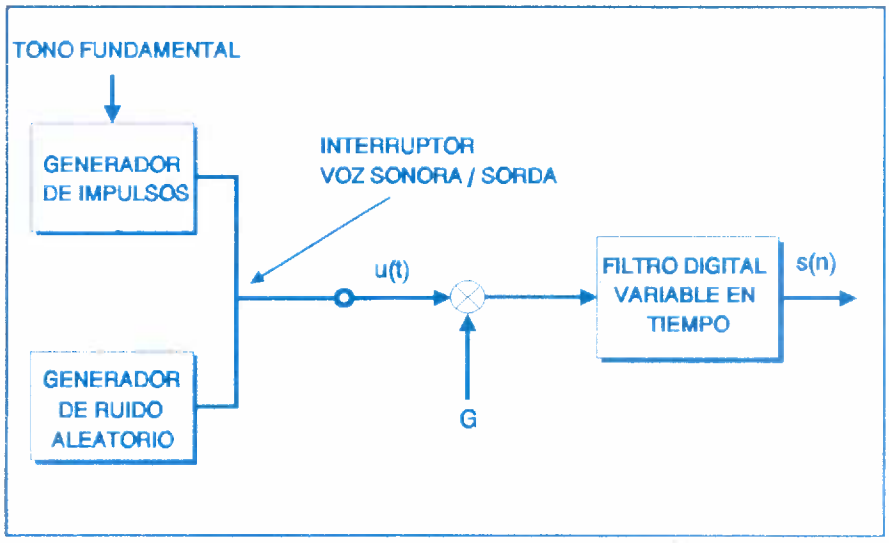

Fig. 6

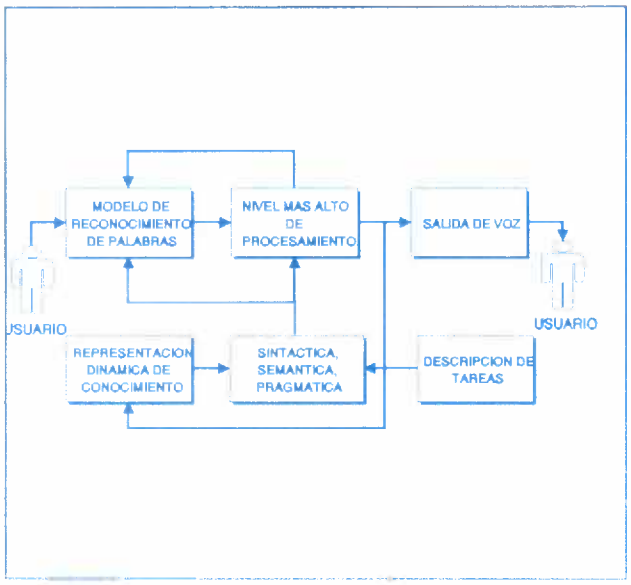

Fig. 7

5. PLANTEAMIENTO DEL MODELO HOMBRE-MAQUINA

El modelo (Fig 7.) comienza cuando un usuario genera una señal de VOZ (habla) para lograr que se ejecute una tarea dada. La entrada hablada (señal del habla) es decodificada en una serie de palabras que son significativas de acuerdo con la sintaxis, semántica, y pragmática de la tarea del reconocimiento.

El significado de las palabras reconocidas es obtenido por un procesador de nivel más alto que utiliza una representación de conocimiento dinámica para modificar la sintaxis, semántica, y pragmática de acuerdo con el contexto de lo qué ha reconocido previamente. La retroalimentación de la caja de procesamiento de más alto nivel reduce la complejidad del modelo de reconocimiento limitando la búsqueda para entradas de sentencias (habla) válidas del usuario.

El sistema de síntesis responde al usuario en forma de una salida de voz, o equivalentemente, en forma a la acción solicitada que es ejecutada cuando el usuario prepara más entradas.

Como no se conoce aún el reto definitivo para el reconocimiento y la generación de Voz por computador, son un campo de investigación con objetivos, métodos y aplicaciones bien definidos, en los que hay mucho trabajo que realizar a distintos niveles "teórico-prácticos" y en distintas materias (Procesamiento digital de señales, Acústica, Fonética, Reconocimiento de formas, Inteligencia Artificial, etc).

6. BIBLIOGRAFIA

(1) Lawrence Rabiner - Biing Hwang Juang, "FUNDAMENTALS OF SPEECH RECOGNITION". Ed. Prentice Hall signal processing series.
(2) Francisco Casacubierta, Enrique Vidal, "RECONOCIMIENTO AUTOMATICO DEL HABLA". Ed. Marcombo boixareu editores.
(3) Creative Labs, "VOICE ASSIST" and "TEXT TO SPEECH" User s Guide.

NOTAS DE PIE DE PAGINA

[1] El **Aparato Fonador**, encargado de generar la Voz a partir de una corriente de aire expulsada de los pulmones. El **Aparato Auditivo**, interpreta las señales de onda que se propagan a través del aire.
[2] Análisis de la información implícita en las formas de onda de la señal vocal.
[3] Sonidos sonoros/sordos, fricación, tono, energía, sonidos nasales, formantes.
[4] Asignación de unidades fonéticas a los diferentes sonidos de la señal vocal.
[5] La segmentación es un proceso de análisis de la onda, encargado de dividirla en partes, dependiendo de las características físicas de la señal acústica.
[6] Algoritmo encargado de incorporar los diferentes patrones de las palabras habladas a una base de datos, en forma de diccionario, conteniendo la palabra en forma de texto y el patrón de la misma.
[7] Proceso de análisis de la señal en el cual el algoritmo se encarga no sólo de dividir la señal en partes, sino también en asignarle un posible rótulo (Fonema) que más caracterice a cada segmento.
[8] Formas de onda de las señales acústicas para cada uno de los fonemas.