

Sistemas de Capítulo 1. reconocimiento y síntesis de voz



← índice resumen figuras tablas introducción **1** 2 3 4 5 6 A B C D E F referencias →

Durante los últimos años ha surgido un nuevo tipo de interfaces humano-computadora que combinan varias tecnologías del lenguaje para permitir el acceso y transferencia de información a través del habla. Estas interfaces basadas en voz involucran principalmente dos tecnologías: reconocimiento y síntesis de voz. El reconocimiento de voz es el proceso de transformar una señal a texto; y síntesis de voz o Tts (Text to speech) el proceso de transformar el texto a una secuencia de sonidos. A continuación se explicará cada una de estas tecnologías.

1.1 Sistemas de reconocimiento de voz

1.2 Sistemas de síntesis de voz

1.1 Sistemas de reconocimiento de voz

Para la creación del diccionario el reconocimiento de voz es una parte importante, ya que es necesario reconocer la palabra que el usuario desea consultar, para después buscarla en la base de datos y dar como resultado la palabra traducida en inglés. El reconocimiento de voz es el proceso de transformar una secuencia de palabras a texto. Los reconocedores se pueden clasificar en:

1. *Reconocedor de propósito específico* , el vocabulario es de dominio restringido; por ejemplo: reconocedor de dígitos
2. *Reconocedor de propósito general* , el vocabulario es de dominio general, por ejemplo: reconocedor para el Español Mexicano.

En la realización del diccionario fue necesario usar el reconocedor de propósito general creado por el grupo TLATOA, ya que las palabras que contiene el diccionario, por ejemplo, blanco, chocolate, elefante, etc. no pertenecen a un vocabulario en específico como en el caso del reconocedor de dígitos.

En la siguiente subsección se dará una breve introducción de la historia del reconocimiento de voz, desde sus inicios, hasta los últimos años.

1.1.1 Historia del reconocimiento de voz

La historia del reconocimiento de voz empezó en el año de 1870. Alexander Graham Bell quiso desarrollar un dispositivo que capaz de proporcionar la palabra visible para la gente que

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

no escuchara. Bell no tuvo éxito creando este dispositivo, sin embargo, el esfuerzo de esta investigación condujo al desarrollo del teléfono. Más tarde, en los años 30 Tihamer Nemes científico húngaro quiso patentar el desarrollo de una máquina para la transcripción automática de la voz. La petición de Nemes fue negada y a este proyecto lo llamaron poco realista[1].

Fue hasta 1950, 80 años después del intento de Bell, cuando se hizo el primer esfuerzo para crear la primera máquina de reconocimiento de voz. La investigación fue llevada a los laboratorios de AT&T. El sistema tuvo que ser entrenado para reconocer el discurso de cada locutor individualmente, pero una vez especializada la máquina tenía una exactitud de un 99 por ciento de reconocimiento[1].

El primer sistema de reconocimiento de voz fue desarrollado en 1952 sobre una computadora analógica que reconocía dígitos del 0 al 9, este sistema era dependiente del locutor. Los experimentos dieron una exactitud de reconocimiento del 98%. Más tarde, en esa misma época, se creó un sistema que reconocía consonantes y vocales[Dudley,58].

Durante los 60's, los investigadores que trabajaban en el área de reconocimiento de voz empezaron a comprender la complejidad del desarrollo de una verdadera aplicación dentro del reconocimiento de voz, y se comenzaron a realizar aplicaciones con vocabularios pequeños, dependientes del locutor y con palabras de flujo discreto. El flujo discreto es la forma como hablan los locutores, es decir, con pequeñas pausas entre palabras y frases. También, durante 1960, la Universidad de Carnegie Mellon e IBM empezaron una investigación en reconocimiento de voz continuo. El impacto de esta investigación se reflejó hasta después de los años 70's.

Para los 70's, se desarrolló el primer sistema de reconocimiento de voz comercial. Se mejoraron las aplicaciones de los sistemas dependiente del locutor que requerían una entrada discreta y tenía un vocabulario pequeño. Por otra parte la Advanced Research Projects Agency (ARPA) de la Sección americana de Defensa se mostró interesada en la investigación de reconocimiento de voz. ARPA comenzó investigaciones enfocándose al habla continua y usando vocabularios más extensos. También se mejoró la tecnología de reconocimiento para palabras aisladas y continuas. En esta misma época se desarrollaron técnicas para el reconocimiento de voz como *time warping*, modelado probabilístico y el algoritmo de retropropagación [Rabiner, 93].

Durante los 80's el reconocimiento de voz se favoreció por tres factores: el crecimiento de computadoras personales, el apoyo de ARPA y los costos reducidos de aplicaciones comerciales. El mayor interés durante este periodo de tiempo era el desarrollo de vocabularios grandes. En 1985 un vocabulario de 100 palabras era considerado grande. Sin embargo, en 1986 hubo uno de 20,000 palabras. También durante esta época hubo grandes avances tecnológicos, ya que se cambió del enfoque basado en reconocimiento de patrones a métodos de modelado probabilísticos, como los Modelos Ocultos de Markov (HMM).

Para los 90's los costos de las aplicaciones de reconocimiento de voz continuaron decreciendo y los vocabularios extensos comenzaron a ser normales. También las aplicaciones independientes del locutor y de flujo continuo (lo contrario al flujo discreto, es decir, en el habla no hay pausas significantes) comenzaron a ser más comunes.

En esta subsección se ha proporcionado una breve introducción de los sistemas de reconocimiento de voz y su historia. A continuación, se describen las principales

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

características acústicas del habla.

1.1.2 Características acústicas

La señal de voz se presenta en forma continua. Sin embargo, el lenguaje transmitido por el habla está formado por palabras, que a su vez están divididas en fonemas. Los fonemas representan la unidad básica del habla.

El conjunto de fonemas producidos en cualquier lenguaje puede ser caracterizado, de acuerdo a sus propiedades, en: vocales (anterior, central y posterior), diptongos, semivocales y consonantes (nasales, oclusivos, fricativas, africativas, flaps, trill); la explicación de cada una de estas se muestra en la tabla siguiente:

Tabla 1.1 Tabla del conjunto de fonemas para el español hablado en México.

Categoría	Descripción	Clasificación
Vocales	Los sonidos producidos por las vocales se generan cuando el aire pasa por los pulmones a la laringe y después a la boca, no existe ninguna obstrucción audible en ninguna de las vocales.	Anterior: /iy/ y /ey/ Central: /aa/ Posterior /ow/ y /uw/
Diptongos	Son vocales en las que la lengua se está moviendo durante la duración del fonema. Cuando el locutor reduce la duración del conjunto formado por dos vocales y las pronuncia de una sola vez, se forma un diptongo.	/ay/ empieza como una /a/ y termina como una /i/ /oy/ empieza como una /o/ y termina como una /i/
Semivocales	Este grupo tiene similitud con las vocales, es por eso que se les llama así. Se producen como las vocales y los diptongos, pero la lengua en posición muy extrema.	/y/ es como una /i/ extrema /w/ es como una /u/ extrema. /l/ es raro encontrarla.
Fricativas	Sonidos producidos por un cierre parcial de la boca.	Labial: /f/ Alveolar: /s/ Velar: /hx/
Stops o Oclusivos	Son sonidos dinámicos, producidos por un cierre total y después una salida repentina de aire. Se clasifican en voiced y unvoiced (hablado o no-hablado), depende de como estén vibrando las cuerdas	Labial: /b/ y /p/ Alveolar: /d/ y /t/ Velar: /g/ y /k/

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

	vocales.	
Flaps y trill	Los flaps son producidos cuando la lengua cierra por un momento corto el tracto vocal.	/r/ y el trill es una secuencia de flaps /rr/.
Africativos	Empiezan como un oclusivo y terminan como un fricativo.	Ejemplo: /ch/.
Nasales	Se producen cuando se cierra el tracto vocal mientras que baja el volumen del habla, dejando pasar el aire por la nariz.	Labial: /m/ Palatal /ny/ Alveolar: /n/ Velar: /ng/

1.1.3 Representación de la señal de voz

Los sonidos consisten en variaciones en la presión del aire a través del tiempo y a frecuencias que podemos escuchar. Una de las maneras para representar el sonido es a través de una onda (waveform), como el ejemplo que se muestra en la figura 1.1.

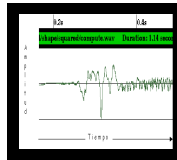


Figura 1.1 Ejemplo de una señal de voz.

Una de las grandes ventajas de éste tipo de gráficas es que no ocupa mucho espacio en memoria. Y una desventaja es que no se describe explícitamente el contenido de la señal en términos de sus propiedades.

Los espectrogramas contienen mayor información sobre los datos de la voz, son una transformación que muestran la distribución de los componentes de frecuencia de la señal. Un ejemplo se muestra en la figura 1.2.

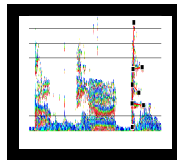


Figura 1.2 Ejemplo de un Espectrograma.

Las partes más oscuras, vistas en la figura, representan la concentración de energía y son denominadas formantes.

Por otra parte es importante mencionar que la capacidad auditiva del ser humano varían en un rango de frecuencias de 20Hz a 20,000Hz. Los sonidos emitidos al hablar se encuentran de

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

100Hz a 15,000Hz en mujeres y en hombres de 400Hz a 15,000Hz. Aunque se cree que la mayoría de la información se concentra debajo de los 8,000Hz [CSLUa].

1.1.4 Arquitectura de un sistema de reconocimiento de voz

Para entender el funcionamiento de un sistema de reconocimiento de voz es necesario conocer sus principales componentes: el extractor de características y el clasificador. Cuando se recibe la señal de voz, ésta pasa por un reconocedor el cual da como resultado la palabra que reconoce. Después hay un procesamiento del lenguaje natural, una representación semántica y finalmente se realiza una acción. La arquitectura para los sistemas de reconocimiento de voz se muestra a continuación:

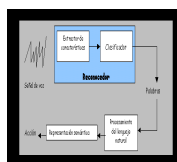


Figura 1.3 Arquitectura de un sistema de reconocimiento de voz.

Como se mencionó anteriormente, en la arquitectura de un sistema de reconocimiento de voz se cuenta con dos procesos importantes en la fase de reconocimiento, estos son los siguientes:

Extracción de características: Los pasos a realizar en este módulo son los siguientes:

1. La señal se divide en una colección de segmentos
2. Se aplica alguna técnica de procesamiento de señales para obtener una representación de las características acústicas más distintivas de segmento.
3. Con base en las características obtenidas, se construye un conjunto de vectores que constituyen la entrada al siguiente módulo.

Clasificador probabilístico: En este módulo se realizan los siguientes pasos:

1. Se crea un modelo probabilístico basado en redes neuronales como modelos ocultos de Markov, etc.
2. Con las probabilidades obtenidas se realiza una búsqueda para encontrar la secuencia de segmentos con mayor probabilidad de ser reconocidos.

1.1.5 Tipos de sistemas de reconocimiento de voz que existen

La meta principal en el desarrollo de aplicaciones usando el reconocimiento de voz es crear sistemas que sean capaces de reconocer diferentes voces, ya sean espontáneas o no, y que el habla sea de manera natural. Esto es difícil ya que aún no se cuenta con un sistema que resuelva las dificultades existentes debido a la variabilidad de las características de la señal acústica. Por esta razón los sistemas de reconocimiento se clasifican de acuerdo a las siguientes restricciones:

A. Dependencia vs. independencia del locutor.

En un sistema dependiente del locutor es más probable que el porcentaje de reconocimiento sea mayor, ya que en este tipo de sistemas, sólo se reconoce a una persona en particular. Las muestras que va a tener el reconocedor sólo pertenecerán a

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

un locutor, lo cual hace más fácil que reconozca el vocabulario.

Cuando es independiente del locutor se trata de un sistema que puede reconocer la voz de cualquier persona, y no importa que ésta no haya sido incluida en el conjunto de entrenamiento. Esta clase de sistemas tiene cierta dificultad, ya que las representaciones paramétricas de la voz son altamente dependientes del locutor. En esta restricción el error es de 3 a 5 veces mayor que en las dependientes del locutor.

B. Palabras aisladas vs. habla continua

El habla continua es cuando un locutor habla de forma natural, con las pausas que son necesarias al hablar. En el habla continua aumenta el grado de dificultad, ya que no es fácil de identificar los límites de las palabras (inicio y final) y pueden estar muy juntas que se confunden con una sola. Sin embargo, cuando son palabras aisladas, el locutor habla más lentamente, esto va a dar como resultado que la probabilidad de reconocimiento sea mayor.

C. Tamaño del vocabulario

Otro factor muy importante es el tamaño del vocabulario, ya que a medida que va creciendo el vocabulario también va aumentando la dificultad para reconocer o van surgiendo nuevos problemas, por ejemplo cuando se confunde una palabra por otra, cuando tarda más en reconocer, etc. Otro factor es la similitud entre palabras, esto es cuando tenemos un vocabulario grande puede haber palabras que se parezcan y esto aumentar la dificultad de reconocimiento. La búsqueda que se lleva a cabo para encontrar la palabra tiene mayor probabilidad de equivocar.

D. Variabilidad y ruido

El ruido puede degradar bastante el nivel de desempeño del reconocedor, y éste es producido por: el ruido ambiental, suspiros, música, etc. Otros factores son es el estado de ánimo del locutor, el ruido producido por el locutor, la calidad del micrófono, entre otros.

Se ha visto que para obtener buen porcentaje de reconocimiento, es necesario tomar en cuenta los factores a los que está expuesto.

1.1.6 Algunas aplicaciones del reconocimiento de voz

Actualmente existen varias aplicaciones en donde se usa el reconocimiento de voz. Un ejemplo es CONMAT (Sistema de conmutador automático) usado en la Universidad de las Américas; este conmutador está en servicio todo el día, lo cual reduce el trabajo de las operadoras, y por lo tanto se cuenta con un mejor servicio en la universidad. Otro sistema que también esta funcionando en la Universidad de las Américas es INFOUDLA, éste realiza consultas del estado de cuenta de los estudiantes, la información que recibe es el ID del estudiante y su NIP. Estas dos aplicaciones son de mucha utilidad para instituciones donde se manejan grandes volúmenes de información y se realizan consultas muy diversas.

Otras aplicaciones del reconocimiento de voz son los siguientes:

- ◆ Ejecución de comandos.
- ◆ Dictado automático
- ◆ Llenado de formas
- ◆ Acceso a información de bases de datos
- ◆ Directorio telefónico automático
- ◆ Servicios financieros por teléfono

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

- ◆ Llamadas por cobrar automáticas.

El uso de sistemas de reconocimiento de voz tiene sus ventajas, como las que se mencionan a continuación:

- ◆ *Eficiencia del habla:* El habla es rápida, flexible y la forma más natural de comunicación.
- ◆ *Cuando es el único canal disponible:* Por ejemplo, no hay teclado, mouse, o pantalla.
- ◆ *El usuario no puede ver lo que escribe:* Puede que tenga las manos o los ojos ocupados y no puede hacer uso de la pantalla, o bien es discapacitado.

1.2 Sistemas de síntesis de voz

La síntesis de voz es el proceso de transformar el texto a sonido (TtS), esto nos sirve para la creación de voz artificial, dadas las palabras escritas, el sintetizador se encarga de pronunciarlas. Las herramientas del CSLU Toolkit contienen sintetizadores en varios idiomas por ejemplo, inglés español mexicano, francés, alemán, entre otros.

1.2.1 Arquitectura general de un TtS

Actualmente existen diversos laboratorios que se dedican a realizar investigaciones y proyectos relacionados con el TtS. En los últimos años se han desarrollado muchos sistemas en diferentes idiomas, con diferente complejidad y diferentes resultados. La arquitectura del TtS es la siguiente:

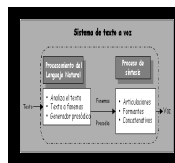


Figura 1.4 Arquitectura de un Sintetizador.

1.2.1.1 Bloque de Procesamiento del Lenguaje Natural

Este bloque se encarga de producir una transcripción fonética del texto leído, además de la entonación y el ritmo deseado para la voz de salida. En la siguiente figura se muestra con detalle cómo se realiza este proceso.

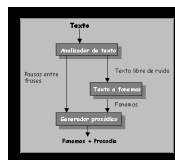


Figura 1.5 Bloque del Procesamiento del Lenguaje Natural (NLP).

El texto de entrada consiste en una colección de símbolos que deben interpretarse como palabras con el fin de tener la idea de lo que se ha escrito. Cada uno de los bloques del NLP que se muestran en la figura 1.5 son las partes esenciales para realizar éste proceso, la

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

explicación de cada bloque se muestra a continuación:

a. Analizador de texto

La función principal del analizador de texto es tomar como entrada cualquier texto y darle el formato adecuado para que sea entendible por el siguiente módulo. El formato que se desea obtener de este proceso es una secuencia de palabras libres de ruido. Además, el analizador de texto realiza la asignación de pausas entre frases.

b. Convertidor de texto a fonemas

Después del análisis de texto, el siguiente paso es convertir de texto a fonemas. Una vez que tenemos las palabras, es necesario asignar su pronunciación para poder generar la señal de voz. Estas pausas se usan para organizar las palabras en un número adecuado de frases.

La transcripción fonética es la asociación de una palabra con los fonemas que lo comprenden. Algunas transcripciones fonéticas se muestran en la tabla 1.2

Tabla 1.2 Ejemplos de transcripciones fonéticas del español.

Palabra	Transcripción fonética
español	/e/ /s/ /p/ /a/ /ny/ /o/ /l/
pastel	/p/ /a/ /s/ /t/ /e/ /l/
cinco	/s/ /i/ /N/ /kc/ /k/ /o/
ardilla	/a/ /r/ /d/ /i/ /dZc/ /dZ/ /a/
blanco	/bc/ /b/ /l/ /a/ /N/ /kc/ /k/ /o/

La idea básica de un sistema Tts consiste en convertir un texto de entrada en una secuencia de fonemas correspondientes a la transcripción fonética de ese texto.

c. **Generador prosódico** Una de las metas principales en el proceso de Tts es la producción de voz con la mayor naturalidad posible [López, 95]. El generador prosódico se encarga de asignar la duración correcta a cada uno de los fonemas, así como una entonación adecuada. Este proceso se puede dividir en dos pasos:

1. Predicción de las duraciones para los fonema.
2. Predicción de la frecuencia fundamental.

1.2.1.2 Bloque del proceso de síntesis

Este bloque se encarga de transformar la información simbólica que recibe el NLP en una voz de salida.

El texto de entrada consiste en una colección de símbolos que deben interpretarse como palabras. Estos símbolos no siempre se asocian directamente con las palabras que representan.

Capítulo 1. Sistemas de reconocimiento y síntesis de voz

Otras veces representan elementos lingüísticos diferentes a las palabras, como en el caso de los signos de la puntuación que pueden influir en la pronunciación de las palabras [Barbosa, 97].

1. Sintetizadores articulatorios

Este tipo de sintetizadores son modelos físicos basados en descripciones detalladas de mecanismos fisiológicos de producción de voz y generación de sonidos en el aparato vocal [Donovan, 96]. Los sintetizadores articulatorios utilizan parámetros tales como el tamaño de la cavidad oral, la tráquea, la posición de la lengua, entre otras variables. Todos estos factores son relacionados para producir la voz.

2. Sintetizadores paramétrica

En este tipo de sintetizadores se genera la voz variando parámetros que aplican señales armónicas. Al modificar los parámetros involucrados en el modelo, se producen sonidos semejantes a los del habla. Para lograr la resonancia del tracto vocal, se incluyen los filtros.[Klatt, 87].

3. Síntesis concatenativa

Aunque la aparición de estos sintetizadores no es reciente, actualmente son los que ofrecen mejores resultados[Barbosa, 97]. Este tipo de sintetizador, forma la voz sintetizada por medio de la concatenación de segmentos de voz (fonemas, sílabas, palabras) previamente almacenados en una base de datos. La base de datos contiene grabaciones hechas de algún locutor.

En este caso se utilizó la síntesis paramétrica, ya que es la más adecuada para la utilización del diccionario por las características que éste tipo de sintetizado muestra.

En este capítulo se describieron los aspectos básicos de los sistemas de reconocimiento y síntesis de voz, que serán la base para la realización del prototipo. En el siguiente capítulo se dará a conocer las herramientas CSLU Toolkit con la que se desarrolló el diccionario.