Anexo B. Tecnologías y herramientas de voz

Los términos “herramientas de voz” y “tecnologías de voz” se han estado usando a lo largo del documento de manera constante, en esta sección se clarifica el significado de estas expresiones y se presentan de manera muy general sus bases teóricas relacionadas.

De inicio, con el término “tecnología de voz”, se hace referencia a un tipo especial de procesamiento computacional de la voz, y con el término “herramienta de voz” se hace referencia a un sistema computacional basado en una o varias tecnologías de voz. Tomando en cuenta esto, el módulo de voz de VELOAT es una herramienta de voz basada en la tecnología de grabación de voz.

Principales tecnologías de voz

En primer lugar es necesario saber que el procesamiento de voz por medios tecnológicos tiene varias sub-disciplinas o ramas de especialización. Las siguientes son las principales.

Grabación de voz

Este primer tipo de tecnología de voz es el más sencillo de manejar. Consta de grabar voz en archivos de audio, mediante software especializado y micrófono. Otras tecnologías más complejas usan o se pueden basar en esta tecnología; un ejemplo es la síntesis de voz.

Síntesis de voz

La síntesis de voz, también conocida como *Text-To-Speech* (TTS), tiene como objetivo la comunicación en el sentido de computadora a seres humanos. Es el proceso de generar voz como salida a partir de texto proporcionado como entrada y de un conjunto de sonidos previamente grabados que se combinan para producir el resultado deseado [Ince, 1992].

Reconocimiento de voz

El reconocimiento de voz tiene como objetivo la comunicación en el sentido de humano a computadora. Es el proceso mediante el cual se analiza una entrada de voz u onda sonora y se trata de obtener el mensaje que contiene, para traducirlo a otro tipo de dato como texto [Ince, 1992]. De todos los tipos de procesamiento de voz, éste es uno de los más complejos debido al gran número de factores que afectan los resultados. Para obtener el significado del texto reconocido, se emplea la tecnología de procesamiento de lenguaje natural.

Procesamiento de lenguaje natural

El procesamiento de lenguaje natural va más allá de convertir entre texto y voz, su objetivo es tratar con los significados formados por conjuntos de palabras o señales acústicas. La generación de lenguaje natural, cuya abreviatura en inglés es NLG, es el proceso de construir salidas de lenguaje natural a partir de entradas no lingüísticas. Generalmente es visto como la transformación de significado en texto. El proceso inverso es la conversión de texto en significado, y es llamado entendimiento de lenguaje natural, cuya abreviatura es NLU [Jurafsky *et al.*, 2000]. Éste también es uno de los tipos de procesamiento más complicados debido a la intervención de la semántica.

Interacción de tecnologías de voz con otros sistemas

Los componentes principales de los sistemas interactivos basados en tecnologías de voz pueden ser uno o varios de los siguientes [Weddon *et al.*, 1990]:

Entrada - reconocimiento de voz

Salida - Síntesis de voz (TTS)

Procesamiento - Análisis y Codificación de voz

Todo lo anterior puede contar con el soporte de tecnologías más complejas, como procesamiento de lenguaje natural e inteligencia artificial, y de estándares relacionados como *Voice XML*. En la figura B-1 se muestra una arquitectura en la que se muestran todas las posibles interacciones de un sistema con tecnologías de voz.

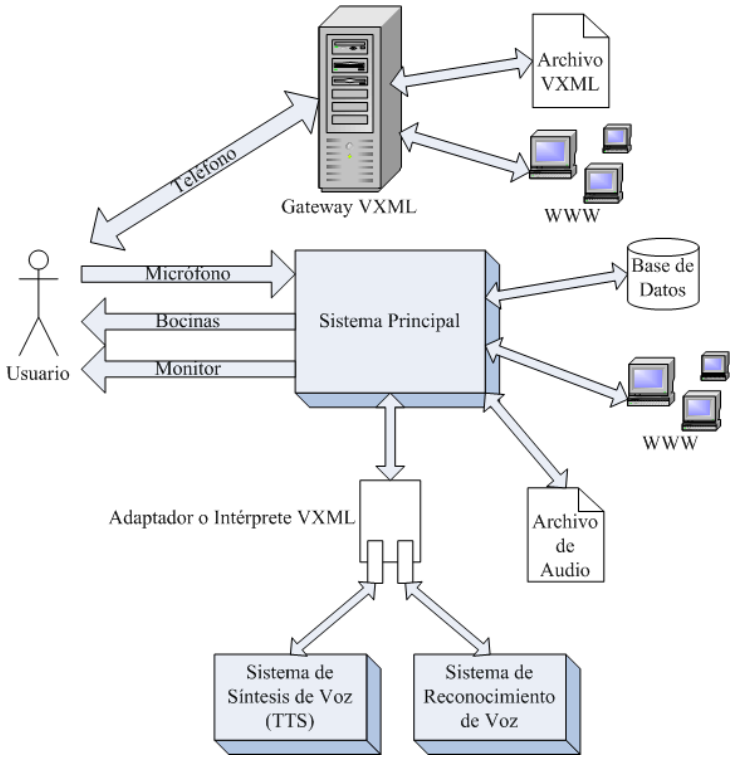

Figura B-1. Arquitectura de sistemas que interactúan con herramientas de voz

Organizaciones dedicadas a la investigación en tecnologías de voz

Existen centros de investigación y universidades en todo el mundo que llevan a cabo investigaciones sobre procesamiento de voz. A continuación se presentan tres ejemplos, que son los más cercanos al desarrollo de este proyecto.

TLATOA

TLATOA es el Grupo de Investigación en Tecnologías del Habla de la Universidad de las Américas Puebla. Este grupo se enfoca en el desarrollo de sistemas de procesamiento de voz para español mexicano y colabora estrechamente con el *Center for Spoken Language Research*, de la Universidad de Colorado (CSLR). Los miembros de este laboratorio son un tipo de usuarios potenciales del sistema desarrollado en este proyecto de tesis. Su sitio Web está en la siguiente dirección: http://info.pue.udlap.mx/~sistemas/tlatoa/

CSLU

El CSLU es el Centro para el Entendimiento del Lenguaje Hablado (*Centrer for Spoken Language Understanding*) de la Escuela de Ciencias e Ingeniería de la Universidad de Salud y Ciencia de Oregon (*Health & Science University of Oregon*). Trabaja en reconocimiento y síntesis de voz, así como en procesamiento de señales. Este centro desarrolló un conjunto de herramientas para el aprendizaje y la investigación de tecnologías de voz, que se llama *CSLU Toolkit*. El TLATOA en la Universidad de las Américas Puebla conoce y usa estas herramientas. El sitio Web del CSLU es el siguiente: http://cslu.cse.ogi.edu/

CSLR

El CSLR es el Centro para la Investigación del Lenguaje Hablado (*Center for Spoken Language Research*) de la Universidad de Colorado. Trabaja en sistemas de diálogo o conversación y en procesamiento de voz encaminados a la creación de recursos educativos y al acceso universal a la información. Su sitio Web está en la siguiente dirección: http://cslr.colorado.edu/

Tecnologías de soporte para el desarrollo de herramientas de voz

Microsoft Speech API & Agents

MS SAPI & Agents son las tecnologías de Microsoft usadas para el desarrollo de herramientas de voz. Ambas tecnologías se pueden obtener desde el sitio de descargas de Microsoft, pero se necesitan conseguir licencias para su uso. MS SAPI es un *kit* de desarrollo de software que implementa funcionalidades de síntesis y reconocimiento de voz, y su documentación se encuentra en:

http://msdn.microsoft.com/library/default.asp?url=/library/en-us/SAPI51sr/html/Welcome.asp

Microsoft Agent es un conjunto de servicios de software que permite incorporar personajes interactivos animados en aplicaciones *standalone* y páginas Web. Estos personajes pueden tener capacidades de habla o de aceptar entrada de voz a través de micrófonos si se usa la tecnología de *Microsoft Agent* en conjunto con Microsoft SAPI. La documentación de *Microsoft Agent* se encuentra en:

http://msdn.microsoft.com/library/default.asp?url=/library/en-us/SAPI51sr/html/Welcome.asp

Voice XML

Voice Extensible Markup Language (*Voice XML*) es una recomendación del consorcio W3C diseñada para la creación de diálogos de audio que incluyan soporte para síntesis de voz, audio digital y reconocimiento de voz, entre otras cosas. Contiene especificaciones para el manejo de entradas de usuario, salidas de sistemas interactivos de voz y recursos externos. Estas especificaciones se pueden acceder a través del sitio Web de esta recomendación, que se encuentra en: http://www.w3.org/TR/voicexml20/

CSLU Toolkit

Como se mencionó anteriormente, *CSLU Toolkit* es un conjunto de herramientas creado por el CSLU de la Universidad de Colorado, para ser usado en instituciones educativas, centros de investigación y distintas industrias que requieran soporte de tecnologías de voz. Es un sistema que integra entre otras cosas herramientas de audio, animación, síntesis y reconocimiento de voz. La documentación de esta herramienta y su versión más reciente se pueden obtener de la siguiente dirección: http://cslu.cse.ogi.edu/toolkit/

Java Speech API y sus implementaciones

Java Speech API es una especificación para el manejo de tecnologías de voz en aplicaciones basadas en Java. Es una interfaz que soporta entre otras cosas reconocimiento y síntesis de voz. La documentación relacionada con esta especificación puede encontrarse en la siguiente dirección:

http://java.sun.com/products/java-media/speech/

Esta especificación tiene distintas implementaciones realizadas por compañías distintas a Sun Microsystems, entre las cuales se encuentran las dos siguientes:

1. Festival

   Festival es un *framework* para el desarrollo de sistemas de síntesis de voz creado por el Centro de para la investigación de tecnología de voz de la Universidad de Edimburgo. Este *framework* soporta síntesis de voz en español e inglés. Está desarrollado en C++, pero tiene una interfaz para el API de *Java Speech*. Su página Web es: http://www.cstr.ed.ac.uk/projects/festival/

2. Free TTS

   FreeTTS es un sistema de síntesis de voz derivado de Festival y basado en otros proyectos de síntesis de voz desarrollados por la *Carnegie Mellon University*. Está desarrollado en Java, y soporta síntesis de voz únicamente en inglés. Su página Web es la siguiente: http://freetts.sourceforge.net/docs/index.php

Java Media Framework

Se trata de un paquete adicional que se extiende de la plataforma de Java 2 (J2SE) y que proporciona soporte para manejar medios basados en tiempo, como audio y video, en aplicaciones Java y *applets*. En el caso del audio, puede manejar los formatos más comunes, incluyendo WAV y MP3. Esta extensión, así como su documentación y más información se pueden obtener de la siguiente página Web:

http://java.sun.com/products/java-media/jmf/