Ejecución de comandos de voz mediante Web Speech API

Javier Pérez[1], Javier Diaz[2], Ivana Harari[2]

[1] Facultad de Informática – Universidad Nacional de La Plata (UNLP)
javi.pzv@gmail.com
[2] LINTI – Facultad de Informática – Universidad Nacional de La Plata (UNLP)
jdiaz@unlp.edu.ar, iharari@info.unlp.edu.ar

Abstract. The predominance of exclusively manual interaction required when interacting with websites is a restrictive condition. This limitation is taken as a starting point and shapes the main motivation for the development of web applications driven by voice commands. It is described and provided a process useful to allow the user to interact with web applications invoking voice commands.

Keywords: Speech recognition, voice commands, voice interfaces, web applications.

1 Introducción

Una interfaz de comando de voz propone un medio de interacción oral que permite a las personas comunicarse con los dispositivos de forma natural utilizando exclusivamente la voz.

Al contrario de los otros mecanismos de interacción, donde las personas deben adaptarse al uso de dispositivos artificiales, en la interacción por medio de la voz se invierten los roles; es la computadora quien se ocupa de entender al ser humano.

Al hacer la analogía entre una interfaz con reconocimiento de voz y el teclado convencional de una computadora, se observa que en el primero la persona accede a un teclado virtual compuesto de miles de teclas de acceso directo, donde cada una de ellas representa una acción a ejecutar a diferencia del teclado que requiere presionar una secuencia de teclas establecidas.

Las ventajas de interactuar a través del uso de la voz producen un incremento significativo en la calidad de comunicación que se da entre el dispositivo y la persona, al mismo tiempo que reduce la cantidad de componentes visuales que deben ser exhibidos para lograr una buena experiencia de usuario.

En la actualidad los sistemas web pueden integrar servicios de reconocimiento de voz externos a la aplicación y de esta forma brindar un medio alternativo de interacción.

A continuación, se describe el sistema de reconocimiento utilizando en el desarrollo de la aplicación "Handsfree for Web" correspondiente al trabajo final de grado "Navegación web dirigida por comandos de voz" [2].

2 Servicio de reconocimiento de voz

Actualmente navegadores de uso masivo como Google Chrome, Firefox y Opera [1] implementan la denominada Web Speech API.

La API de reconocimiento de voz tiene como objetivo dar servicios de análisis y síntesis del habla [2]. Esto permite a los usuarios integrar un sistema de reconocimiento de voz en aplicaciones web. Regularmente estas aplicaciones utilizan el procesamiento de voz para transformar un discurso oral en texto o viceversa y de esta forma proveerle al usuario nuevos mecanismos de interacción.

La API de reconocimiento de voz provista por los navegadores que la soportan cuentan con las siguientes ventajas que facilitan la integración con sistemas web [3]:

- Servicio de procesamiento de voz externo a la aplicación
- Procesamiento remoto
- Procesamiento en tiempo real
- Obtención de resultados parciales durante el proceso de reconocimiento
- Uso gratuito y sin límites
- Sistema de reconocimiento independiente del usuario, de propósito general y capaz de procesar frases con palabras conectadas
- Soporte de múltiples idiomas y dialectos
- Mejora continua del servicio
- Posibilidad de definir un servicio de reconocimiento alternativo al que viene asignado por defecto

Fig. 1. Soporte de API de Reconocimiento de Voz por parte de los navegadores más populares [1].

3 Proceso de ejecución de un comando de voz

Se define un comando de voz como un par (*nombre*, *acción*) donde *nombre* es una frase que puede ser pronunciada por el usuario y *acción* es una tarea que será ejecutada por el sistema una vez que el nombre asociado sea invocado.

En esta sección se describen los componentes y las etapas que conforman el proceso de ejecución de comandos de voz que desempeña la herramienta Handsfree for Web, mientras los usuarios interactúan oralmente [2].

3.1 Captura y envío de señal sonora

Cuando una aplicación web solicita el inicio de los servicios de reconocimiento de voz por primera vez, el navegador solicita al usuario permisos de captura de sonidos a través del micrófono.

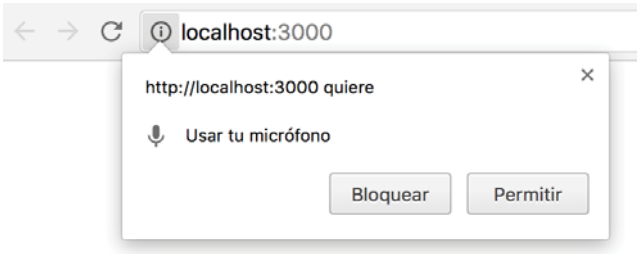

Fig. 2. Permisos solicitados por el navegador Google Chrome al iniciar el servicio de reconocimiento de voz.

La API de Reconocimiento de Voz propone dos formas distintas de captura de sonido.

Reconocimiento de voz continuo

Cuando este modo es seleccionado, el cliente de reconocimiento de voz, captura y procesa el sonido de forma continua hasta que el servicio de reconocimiento es detenido de forma explícita.

Reconocimiento de voz interrumpido

A diferencia del modo continuo, cuando se opta por la modalidad de reconocimiento de voz interrumpido, el servicio finaliza automáticamente una vez que el sistema detecta una pausa en el sonido capturado.

3.2 Recepción de transcripción

Mientras el flujo de datos correspondientes al sonido capturado es enviado activamente, se reciben transcripciones parciales y no finales correspondientes al sonido procesado.

La herramienta puede recibir, como resultado del proceso de reconocimiento del sonido, hasta diez transcripciones posibles, cada una de ellas con un grado de precisión y con una marca que indica si son resultados finales o parciales.

3.3 Inferencia y resolución de un comando de voz

Las transcripciones parciales recibidas pueden ser utilizadas para dar feedback al usuario acerca del estado del proceso de reconocimiento de voz. A la hora de determinar un comando de voz, solo debe tenerse en cuenta las transcripciones de tipo final que superen cierto nivel de precisión.

A pesar de que la Especificación de Web Speech API describe la posibilidad de definir reglas gramáticas [2], actualmente las implementaciones del servicio no soportan la incorporación de las mismas, cuando son definidas por el cliente. Esto provoca en que las transcripciones recibidas resultantes del proceso de reconocimiento corresponden a texto que no obedece ninguna regla gramatical asignada por la aplicación web.

Regularmente se obtienen múltiples transcripciones que pueden ser comandos de voz y no coinciden con exactitud con las acciones disponibles para ser ejecutadas. Por ejemplo, al indicar verbalmente el comando de voz “abrir” se puede obtener la transcripción “habría”. Al buscar por el comando “habría” se encuentra que el mismo no existe. Situaciones de este tipo dan origen en la necesidad de realizar aproximaciones con el fin de determinar si existe un comando similar al mencionado que satisfaga la intención del usuario.

Con la finalidad de realizar aproximaciones precisas, se propone en principio transformar todos los comandos disponibles y las transcripciones recibidas en fonemas utilizando el algoritmo Metaphone [4]. Posteriormente, utilizando el Coeficiente de Sorensen-Dice [5], se comparan los fonemas correspondientes a las transcripciones y los comandos disponibles con el propósito de obtener el par (transcripción, nombre de comando) que presenten mayor grado de similitud.

La obtención del par (transcripción, nombre de comando) con mayor similitud no implica que el mismo sea un buen resultado. Puede ser que el grado de similitud obtenido sea bajo y por consecuencia lejos de representar la intención del usuario.

Luego de sucesivas pruebas realizadas a través de ensayo y error utilizando la herramienta Handsfree for Web [6], se concluyó que el par (transcripción, nombre de comando) debe poseer al menos un 70% de nivel de similitud para que el comando sea considerado como representativo de las intenciones del usuario.

La etapa de resolución de comando de voz puede finalizar con dos resultados; la existencia de un comando a ejecutar o la ausencia del mismo. En caso de no existir un comando a ejecutar se puede informar por pantalla que no existe el comando de voz indicado y/o solicitar que el mismo sea pronunciado nuevamente. De lo contrario se procede a la etapa siguiente.

3.4 Ejecución de un comando de voz

Una vez elegido el comando de voz, la reproducción del mismo consiste simplemente en ejecutar el código definido para el mismo.

4 Resultados obtenidos

Durante la fase de evaluación de la aplicación Handsfree for Web se realizaron cinco entrevistas a personas pertenecientes a distintos rangos etarios, idioma y género. A las mismas se les solicitó que naveguen la web invocando comandos de voz [2].

	Maitena	Alicia	Carlos	Belén	Ivan
Idioma	Español	Español	Español	Español	Inglés
Comandos Invocados	56	92	30	38	94
Comandos interpretados correctamente	51	83	27	33	80
Comandos interpretados incorrectamente	5	9	3	5	15
Tasa de acierto	91%	90%	90%	86%	85%

Fig. 3. Resultados obtenidos durante las pruebas de usabilidad correspondientes a la aplicación Handsfree for Web

Los resultados obtenidos durante las evaluaciones realizadas indican una tasa de acierto promedio cercana al 90%. Al analizar los comandos de voz fallidos, se observó que la mayoría de ellos fueron producidos por la incapacidad de reconocer la palabra "click" cuando la aplicación estaba configurada en modo español. Una situación similar sucedió con los comandos de voz correspondientes a números. Al momento de realizar la evaluación, el software de reconocimiento brindaba la representación literal de los números, en vez de la numérica. Por ejemplo, resolvía la palabra "quince" en vez de el número "15". Luego de ajustar la aplicación para que sea capaz de procesar representaciones literales de números y agregar un alias al comando "click" llamado "presionar", se obtuvieron tasas de acierto cercanas al 100% para los idiomas español e inglés.

A continuación se brindan mecanismos adicionales mediante los cuales se puede incrementar la tasa de acierto correspondiente al reconocimiento de comandos de voz.

5 Mejoras al reconocimiento de comandos de voz

En la sección anterior se describió una forma simplificada en la cual se puede inferir un comando de voz a partir del texto resultante del análisis de la expresión oral del usuario. Se considera que este proceso puede ser mejorado desde los siguientes aspectos:

5.1 Mejorar la captura del sonido

Desde el punto de vista del hardware el medio convencional por el cual se captura la voz del usuario es el micrófono, el cual puede ser interno o externo. Estos micrófonos, al no ser de uso profesional, poseen deficiencias a la hora de capturar el sonido.

Atendiendo al software, el controlador instalado en el sistema operativo suele aplicar filtros con la intención de disminuir el ruido ambiental. Si bien el contexto en el cual el usuario utiliza la herramienta no es un ambiente controlado y estaría fuera del alcance de este trabajo, se podría proveer al usuario un dispositivo especializado de captura de audio con filtros de sonidos apropiados tales como; balance de graves y agudos, reducción de ruido ambiental, y otras normalizaciones. De esta forma la captura específica de la voz de la persona sería más eficiente y como consecuencia de ello se obtendrían mejores resultados a la hora de realizar el reconocimiento de voz.

5.2 Mejorar el reconocimiento de voz

La API de reconocimiento de voz descripta permite la definición de un servicio de reconocimiento diferente al provisto por defecto. Con el fin de mejorar los resultados del proceso de reconocimiento, se podrían utilizar otros servicios alternativos que existen en el mercado.

En la actualidad existen diversos sistemas de reconocimiento de voz, cada uno de ellos presentan diversas características, las cuales son relevantes a la hora de obtener mejores tiempos de respuesta y tasas de reconocimiento de voz [6].

5.3 Mejorar la inferencia del comando de voz

Una vez obtenida la transcripción resultante del proceso de reconocimiento de voz que dio lugar la expresión verbal del comando por parte del usuario, es necesario determinar en concreto cual es el comando de voz a ejecutar. Por lo general, la identificación del comando no concuerda en un 100% con la transcripción recibida por el servicio de reconocimiento de voz. Por ejemplo, cuando el usuario dice “click”, se suele obtener la transcripción “clip”. Al intentar determinar el comando correspondiente a “clip” se encuentra que el mismo no existe.

Al no contar con sistemas de reconocimiento de voz con tasas de acierto del 100%, es importante analizar los resultados obtenidos, con el fin de inferir cuál fue la intención del usuario, y a partir de ello ejecutar el comando de voz que tenga una mayor probabilidad de acierto.

Como se vio en la sección anterior, es posible realizar comparaciones fonéticas entre los posibles comandos que pueden ser ejecutados y las transcripciones resultantes de la expresión del usuario, con el objetivo de determinar los posibles comandos de voz que él mismo quiso ejecutar. El comando de voz de mayor probabilidad de acierto es el elegido.

Si bien se pueden obtener buenos resultados a la hora de inferir los comandos, el proceso descripto puede ser mejorado. A continuación, se describen algunas posibles mejoras.

Luego de la ejecución de un comando, el usuario podría indicar si el comando ejecutado fue el deseado. De esta forma, se podrían almacenar los casos fallidos de resolución de comandos con el fin de no volver a reproducir el fallo.

Cada vez que una persona menciona un comando de voz desconocido para la aplicación, se le podría mostrar una lista de comandos de voz disponibles y consultarle qué comando de voz deseó invocar. De esta forma, la próxima vez que el usuario invoque las mismas palabras, se contará con la información necesaria para inferir el comando apropiadamente.

La correcta selección de los comandos de voz es un aspecto importante a tener en cuenta. A la hora de inferir los comandos, se obtienen mejores resultados si la pronunciación de los mismos es fonéticamente diferente. Por ejemplo, si analizamos el comando “dos”, correspondiente al número "2", es en cierta forma es similar al comando “favoritos”. En caso de visualizar errores en la resolución de comandos, se podrían renombrar alguno de ellos a “número dos” o “lista de favoritos” respectivamente.

6 Conclusión

El servicio de reconocimiento de voz provisto por los navegadores web posibilita a las aplicaciones web brindar nuevos mecanismos de interacción.

La incapacidad de especificar gramáticas a la hora de inicializar los servicios de la Web Speech API motiva la necesidad de inferir comandos de voz a partir de lo mencionado por el usuario.

Se propuso un método de reconocimiento e inferencia de comandos, el cual realiza comparaciones fonéticas entre las transcripciones textuales resultantes de lo expresado verbalmente por el usuario y las acciones soportadas por el sistema.

El proceso de ejecución de comandos de voz, permite capturar la intención del usuario de una forma efectiva y confiable. Estas características hacen posible su utilización como punto de entrada de una interfaz de voz.

Referencias

1. Speech Recognition API, http://caniuse.com/#feat=speech-recognition
2. Web Speech API Specification, https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html
3. Adorf, J.: Web Speech API. KTH Royal Institute of Technolog, Stockholm (2013)
4. Lawrence Philips' Metaphone Algorithm, http://aspell.net/metaphone
5. Rodriguez-Salazar, M. E., Álvarez-Hernández, S., Bravo-Núñez, E.: Coeficientes de Asociación. Plaza y Valdés (2001)
6. Perez, J.: Navegación web dirigida por comandos de voz. UNLP, La Plata (2017)