



Who Talks: Subtitulación Inclusiva

Proyecto de la Fundación ONCE

Realizado por:
Andrei Erhan
Juan Emilio Crespo Perán

Índice

Introducción	3
Investigaciones realizadas	4
Bluetooth	4
WiFi	7
Aplicaciones speech-to-text	11
Arquitectura	14
Experimentos realizados	15
Reconocimiento de voz	15
Calidad del audio	15
Chat en tiempo real	16
Uso de librería Web Speech API	17
Desarrollo	18
Interfaz	19
Dificultades encontradas	21
Investigación	21
Desarrollo	21
Conclusiones	22

Introducción

Es muy común que las empresas hoy en día realicen una o varias reuniones durante la semana, pero un factor muy a tener en cuenta, es conocer si alguna de las personas que asisten a la reunión, tienen problemas de audición.

Esto es un problema porque es probable que no sean capaces de captar todas las cuestiones que se plantean, y por tanto, no se pueden comunicar de una forma eficiente con las demás personas de la reunión.

La solución que se plantea para resolver este problema, es desarrollar una aplicación que realice una subtítulos automática de la reunión en tiempo real para que todas las personas puedan participar de forma activa.

Para ello, los requisitos que debe tener la aplicación son los siguientes:

- Recoger el audio para transcribirlo a texto.
- Identificar correctamente a la persona que está hablando en ese momento mediante huella de voz.
- Los mensajes transcritos deben mostrarse en tiempo real.
- Tener una buena calidad del subtítulo.

La solución inicial que se ha planteado es utilizar un dispositivo centralizado que se encargará de recoger el audio de los asistentes mediante un micrófono, identificar a las personas que están hablando en ese instante y también de transcribir el audio a texto. Para ello, era necesario que se implemente una base de datos que incorpore las huellas sonoras de cada persona de la reunión, lo cual era muy complejo de implementar.

También se ha comentado la posibilidad de recoger el audio mediante un teléfono móvil y que envíe el audio al dispositivo central mediante Bluetooth, de forma que el dispositivo central se encargue de procesar el audio y de transcribirlo.

Se realizará un trabajo de investigación que se centrará principalmente sobre los siguientes temas:

- Búsqueda de librerías de transcripción de audio y de grabación de voz.
- Separación de audio mediante la identificación de las personas a través de su huella de voz.
- Comunicación de dispositivos mediante Bluetooth.

Investigaciones realizadas

Este capítulo se corresponde con las investigaciones que se han realizado a lo largo del desarrollo del proyecto, en la que además de las investigaciones principales, se incluyen otro tipo de investigaciones relacionadas.

Sobre cada una de las investigaciones, se adjunta los enlaces y una breve descripción de los resultados que se han obtenido.

Bluetooth

El manejo de datos entre dispositivos mediante Bluetooth se realiza mediante las clases BluetoothServerSocket para el dispositivo configurado como servidor, y BluetoothSocket para los dispositivos clientes.

El servidor acepta múltiples conexiones y las guarda en un vector de conexiones para recibir y enviar los mensajes de los clientes en el grupo de reunión.

Los datos que acepta son similares a los datos que aceptan los Sockets de Java que son DataInputStream y DataOutputStream, por lo que el envío de ficheros de audio o de texto no será ningún problema.

La idea principal que se tiene es que, por cada conexión que se reciba en el servidor, se envíe además los datos del usuario para que el servidor no tenga que pedir los datos del usuario constantemente y optimizar los recursos en el envío de datos en las posteriores conexiones, donde únicamente se enviará los ficheros de audio o textos.

Tema	Descripción	Importancia para el proyecto	Enlace
System and method for maintaining and providing personal information in real time	Almacenamiento de información de un determinado usuario a una base de datos en tiempo real.	MEDIA	https://patents.google.com/patent/US7877275B2/en
	Evolución de la red móvil, describiendo las tecnologías que suele usar, estando entre		https://www.researchgate.net/profile/Johan_De_Vriendt

Mobile network evolution: a revolution on the move	ellas Bluetooth que contiene información muy básica y poco relevante.	MUY BAJA	/publication/3196666_Mobile_network_evolution_on_the_move/links/0fcfd50f8154cc8800000000/Mobile-network-evolution-A-revolution-on-the-move.pdf
Real-time speech-to-text conversion in an audio conference session	Método de proporcionar recursos de audio en tiempo real. Para ello tiene que haber un sistema que sea capaz de interactuar con varios dispositivos de computación a través de una determinada red. Cada dispositivo va a generar un flujo de audio conteniendo una señal de audio que se enviará en tiempo real al servidor, el cual se encargará de procesar ese flujo y extraer ese audio.	ALTA	https://patents.google.com/patent/US9560206B2/en
Ricocheting Bluetooth	Documento de PAGO . El propósito de este documento es especificar el potencial que tendría un sistema al integrar Bluetooth. Sería necesario desarrollar ciertos algoritmos de reconocimiento de voz/datos para que diferentes sistemas inalámbricos puedan comunicarse mediante Bluetooth sin problemas.	MEDIA	https://ieeexplore.ieee.org/abstract/document/895713
Audio streaming over Bluetooth: an adaptive ARQ timeout approach	Documento de PAGO . El propósito de este documento es centrarse en enviar audio en streaming por Bluetooth. Para ello, se menciona que se hace una mejora al mecanismo ARQ de la capa de Bluetooth para compensar la degradación del canal y para transmitir mejor la transmisión de audio.	ALTA	https://ieeexplore.ieee.org/abstract/document/1284031
Streaming audio over Bluetooth ACL links	Documento de PAGO . El propósito de este documento es describir la tecnología Bluetooth para transmitir ficheros de audio.	ALTA	https://ieeexplore.ieee.org/abstract/document/1197542

Real-time voice over IP over Bluetooth	Documento de PAGO . El propósito de este documento es describir una nueva tecnología llamada VoIPoB que combina conexión inalámbrica con Internet. La información de audio se enviaría como bits de datos a través del enlace ACL en lugar del SCO, usado hasta entonces. Según un simulador de Matlab, se producen mejores resultados de esta manera.	MUY ALTA	https://ieeexplore.ieee.org/document/4062201
--	---	----------	---

Conclusiones sobre el uso de la tecnología Bluetooth

Después de realizar un trabajo de investigación sobre la tecnología Bluetooth, hemos descartado esta tecnología por varias razones:

- **Compatibilidades entre lenguajes:** El proyecto consta de dos aplicaciones en dos lenguajes distintos, por lo que es bastante probable que existan problemas de compatibilidades entre ellos al utilizar Bluetooth, ya que no se utiliza demasiado.
- **Problemas de conexión y transmisión de datos:** Existen situaciones en las que las comunicaciones a través de Bluetooth pueden causar desconexión y en cuanto al envío de datos, pueden causar demoras o incluso errores en el envío de datos, y como se trata de realizar envíos de datos de forma continua y se quiere conseguir una calidad de subtítulo de al menos de un 90%, esta tecnología no cumple con lo exigido.
- **Streaming de audios:** El soporte para realizar streaming por Bluetooth es bastante complicado, ya que la información acerca de ello es bastante escasa, puesto que hemos encontrado pocos artículos específicos de utilidad. En conclusión, pensamos que realizar streaming de audios por Bluetooth no es una buena elección.
- **Gasto energético:** Dejar habilitado el Bluetooth de un dispositivo móvil o tablet consume bastante batería en el dispositivo, y dado que las reuniones no son breves, esto conlleva a que muchas personas queden desconectadas virtualmente de la reunión.
- **Límite de conexiones entre dispositivos:** Cada ranura Bluetooth admite una cantidad limitada de usuarios de forma simultánea, de forma que si en una reunión hay bastantes personas, se tiene que ampliar el hardware y eso conlleva a más gastos económicos.
- **Velocidad de transmisión variable:** Depende de la distancia que haya entre el móvil o tablet de la persona y el servidor, los datos se enviarán más o menos rápidos. Para dispositivos con Bluetooth 5, la velocidad de transmisión es de 2 MB/S siendo menor en dispositivos que tengan una versión inferior de Bluetooth (<https://www.bluetooth.com/blog/exploring-bluetooth5-whats-new-in-advertising/>). Es

posible que se comprometa el objetivo de que las transcripciones sean en tiempo real.

- **Limitación geográfica:** Los asistentes de la reunión deben estar físicamente en la misma sala, por lo que sí existe una reunión con personas que están físicamente en otro lugar, la aplicación no funcionará para ellos.

WiFi

Las conclusiones alcanzadas después de investigar la tecnología Bluetooth, han sido la de desechar esta tecnología. En su lugar, se ha reemplazado por la tecnología WiFi o **TCP/IP**, es decir, envío de datos a través de Internet.

Tema	Descripción y/o comentarios	Importancia para el proyecto	Enlace
Comparative study of VoIP over WiMax and WiFi. Pag 447	Se hace una comparación entre usar WiFi o WiMax en el caso de transmisión de datos mediante VoIP y se llega a la conclusión de que WiMax es mucho más eficiente que WiFi.	ALTA	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.442.8665&rep=rep1&type=pdf#page=447
Voice activated command and control with speech recognition over WiFi	<p>Hidden Markov Models: método estadístico por excelencia para reconocimiento de voz.</p> <p>El objetivo en este documento es construir un framework con aplicaciones para reconocimiento de voz mediante una red inalámbrica. Para ello usarían Java Speech API, JSML y JSGF, que ofrecen herramientas muy potentes para dicho fin.</p> <p>Se diferencian dos tipos de sistemas de reconocimiento de voz: dependiente de la voz e independiente de la voz (En este proyecto se podría intentar desarrollar el dependiente como otra versión de la aplicación. Para ello, antes de nada, al iniciar la aplicación, que haya una opción para que aprenda sólo tu voz, teniendo que leer varios ejemplos de texto hasta que el software tenga las suficientes características de tu voz y te pueda identificar. Con ello, se evitaría que el micrófono de una determinada persona, que está callada, recoja el audio de la persona más cercana y que salga en la aplicación dos mensajes iguales, uno de ellos identificando erróneamente al</p>	MUY ALTA	https://reader.elsevier.com/reader/sd/pii/S0167642305000882?token=7A60348ADA7A24C9E7372667AE015D50D977B6DCEDC47FD58B46B7125EB5F6280CDB4773E7AD925C6653E9BD9856509A

	<p>interlocutor. Una vez implementada esta versión, compararla con la solución propuesta por nosotros (filtrado por decibelios) y analizar cuál de ellas funciona mejor). Imaginemos que estamos en el caso independiente de la voz. Incluiría una base de datos de los distintos fonemas que existen, un sistema de procesamiento y otro de decodificación de la voz. El sistema de procesamiento analizaría la voz de entrada proporcionada por el micrófono y se extraería las características de las palabras, las cuales se usarán en la parte de decodificación. El sistema de decodificación compararía las características del vector de entrada y las compararía con los de la base de datos generando como resultado una probabilidad. Para ello se usaría el modelo HMM y se generaría un grafo para cada fonema de entrada usando Sphinx4 como decodificador. Cada HMM tiene una transición a varios nodos del grafo, y, usando el algoritmo Viterbi, se encontraría el mejor camino basado en el mejor resultado que consiga. Finalmente, el servidor mostraría por pantalla el fonema que se ha recogido de la entrada en tiempo real.</p>		
WiFi real-time streaming and Bluetooth coexistence	<p>Este documento presenta el caso en el que un dispositivo se comunica con varios dispositivos para intercambiar información mediante WiFi, para que a continuación se use Bluetooth con el objetivo de enviar una determinada información a un periférico, por ejemplo unos cascos. La coexistencia de las dos tecnologías puede generar ciertas colisiones, por lo que el autor explica una arquitectura para reducir la probabilidad de que se produzcan. En nuestro caso no es muy útil puesto que toda la comunicación se produciría por WiFi o WiMax (sí lo sería si se activaran micrófonos a los dispositivos).</p>	MEDIA	https://patents.google.com/patent/US9485778B2/en
Reliable audio-video transmission	<p>En este documento se explica un proyecto de innovación centrado en proporcionar un mayor grado de calidad y confiabilidad del servicio mediante redes como WiFi.</p>	MEDIA	https://patents.google.com/patent/US8184657B2/en

system using multi-media diversity			
An experimental evaluation of Apple Siri and Google Speech Recognition	<p>El artículo consta de un experimento realizado hacia las tecnologías de Siri y GSR. El experimento consiste en estudiar el concepto del lag en transmisiones streamings de audio. Hay dos aspectos fundamentales por las que se produce que son la pérdida de paquetes y el jitter. Se llega a la conclusión de que GSR conlleva menos retardo que Siri por lo que es recomendable seguir su método. Se ha comprobado que si las transmisiones se realizan mediante TCP, aunque se reduce considerablemente la pérdida de paquetes y se mantiene entre los valores de jitter óptimos, se produce un pequeño retraso en lo que tarda el paquete en enviarse y recibirlo de vuelta. En cambio, usando UDP y codificación de red (networking coding) se reduce considerablemente el retraso. Aún así, se espera que se siga mejorando las técnicas para reducir al máximo el retraso.</p>	ALTA	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.706.5567&rep=rep1&type=pdf
Methods of searching using captured portions of digital audio content and additional information separate therefrom and related systems and computer program products	<p>En este artículo lo único útil para este proyecto es el tratamiento del audio en la parte servidora. Una vez que se recibe un bloque de audio en el servidor, éste ejecutaría la herramienta de transcripción a texto. Una vez obtenido el texto estimado, se buscará dicho texto en todas las palabras de un determinado vocabulario contenidas en una base de datos, con el objetivo de ver si el texto que se ha transcrito está incompleto o incorrecto, y en dicho caso, recoger la palabra del vocabulario que más se parezca.</p>	ALTA	https://patents.google.com/patent/US8239480B2/en
Multi-modal content and automatic speech recognition in wireless telecommunication systems	<p>Nada nuevo con respecto a los anteriores artículos. Destacar que en este artículo también se defiende el uso de canales UDP como modo de transferencia de datos de audios.</p>	BAJA	https://patents.google.com/patent/US7382770B2/en

Hybridized client-server speech recognition	Este artículo especifica las dos arquitecturas que pueden haber en el caso de reconocimiento de voz. Una de ellas consiste en que la parte cliente tenga varias funcionalidades y que habría menos peticiones al servidor. La otra es que el servidor haga todo la parte de procesamiento del audio, para ello puede tener un módulo de reconocimiento de voz para cada cliente, lo que puede evitar un posible retardo en el módulo central de reconocimiento de voz (si solo hubiera un módulo de reconocimiento) y repartir la carga en diferentes módulos.	MEDIA	https://patents.google.com/patent/US9674328B2/en
ZigBee WiFi coexistence report	Se describe un experimento que han realizado para detectar posibles interferencias del Wifi o Bluetooth con ZigBee pero no tiene ninguna relación con el propósito de nuestro proyecto, quizá sólo los pasos que han realizado para hacer el experimento.	MUY BAJA	http://vip.gatech.edu/wiki/images/8/8e/Zigbee_WiFi_Coexistence_-_White_Paper_and_Test_Report.pdf
Real-time caption streaming over WiFi network	Documento de PAGO . Falta por investigar		https://ieeexplore.ieee.org/abstract/document/1270629 http://web.cs.ucla.edu/ST/docs/itre2003.pdf
A TDMA-based mechanism to enforce real-time behavior in WiFi networks	Documento de PAGO . Falta por investigar		https://ieeexplore.ieee.org/abstract/document/4638758 https://paginas.fe.up.pt/~pportuga/research/fct_2009/paper_WFCS_2008_WIP.pdf
How can I stream audio from my phone over WiFi to a speaker or a receiver	Falta por investigar		https://android.stackexchange.com/questions/29566/how-can-i-stream-audio-from-my-phone-over-wifi-to-a-speaker-or-receiver
Streaming voice between Android Phones over WiFi	Falta por investigar		https://stackoverflow.com/questions/9238376/streaming-voice-between-android-phones-over-wifi

Las ventajas que se han obtenido al realizar las investigaciones anteriores frente a la tecnología Bluetooth son las siguientes:

- **Sin limitación geográfica:** Los asistentes de la reunión no tienen porqué estar físicamente en el mismo lugar para poder utilizar la aplicación y recibir mensajes transcritos.
- **Ausencia de dispositivos externos:** No existe dependencia de llevar un dispositivo extra a una reunión y sin pagar ningún coste en su fabricación.
- **Sin problemas de desconexión:** Las comunicaciones a través de TCP/IP son bastante estables y no hay límite de conexiones simultáneas entre usuarios.
- **Velocidad de transmisión:** Si se utilizan datos móviles la velocidad de transmisión a través de 4G es de 12'5 MB/s como mínimo (https://es.wikipedia.org/wiki/Telefon%C3%ADa_m%C3%B3vil_4G) y si se conectan por Wifi con Adsl de 24Mbps se obtiene una velocidad de transmisión de 3 MB/s(<https://www.adslzone.net/2016/10/18/limite-velocidad-del-cable-cobre/>). En ambos casos, las velocidades de transmisión son siempre superiores a las ofrecidas por Bluetooth, por lo que se puede garantizar el tiempo real sin tener en cuenta que hoy en día, se utiliza la fibra óptica que ofrece una mayor velocidad.
- **Compatibilidad entre distintas plataformas:** El uso del protocolo TCP/IP está bastante extendido y está soportado por todos los lenguajes, por lo que las comunicaciones entre distintos lenguajes serán más sencillas de implementar. Además, si se crean tanto aplicaciones de escritorio como móviles, el desarrollarlas llevará menos tiempo debido al gran soporte que existe.
- **Número de conexiones ilimitadas:** La aplicación alojada en un servidor puede aceptar multitud de peticiones de usuarios, sin embargo, con la tecnología Bluetooth se permite hasta un máximo de 4 usuarios.

Por otro lado, como única desventaja frente a la tecnología Bluetooth, es que se debe contratar un servidor externo para alojar la aplicación y que puedan los clientes conectarse, pero se obtienen más beneficios si analizamos las ventajas que se obtienen frente a la tecnología Bluetooth.

Aplicaciones speech-to-text

Las investigaciones realizadas sobre librerías que realicen el tratamiento de audio a texto son las siguientes.

Tema	Problema / Solución	Enlace
Detección de silencio en tiempo real	Recoge audio cada X milisegundos pero no realiza la detección de silencio.	https://stackoverflow.com/questions/43338528/android-real-time-silence-detection
Librería para pasar de audio a texto con posibilidad de detectar si una persona ha dejado de hablar	Testeo de la aplicación, fallos de compilación difíciles de resolver	https://github.com/Kaljurand/speechutils
Detección de silencio	La mayoría de audios que produce están corruptos y	https://stackoverflow.com/questions/19145213/android-

	no se pueden reproducir.	audio-capture-silence-detection/19752120
Procesamiento de audio en tiempo real con AudioRecord	Testeando la aplicación hemos visto que no funciona muy bien, fallos al coger el audio.	https://developer.android.com/reference/android/media/AudioRecord https://github.com/ashok1995/real-time-audio-processing-using-AudioRecord-android/blob/master/MainActivity.java
Android: need to record mic input	No hemos conseguido que funcione pero creemos que podría funcionar o coger algunos aspectos importantes de allí.	https://stackoverflow.com/questions/6959930/android-need-to-record-mic-input
Grabadora	<p>Problema al grabar un audio en fichero.</p> <p>Este problema ha quedado resuelto debido a que se utilizaba un método para obtener la ruta de la carpeta externa y que estaba en desuso por una versión de API inferior a la desarrollada.</p>	https://www.youtube.com/watch?v=M3NvEen7RyE https://developer.android.com/reference/android/os/Environment https://stackoverflow.com/questions/17540737/java-io-filenotfoundexception-open-failed-eaccess-permission-denied-on-device?rq=1 https://stackoverflow.com/questions/44808923/java-io-filenotfoundexception-permission-denied-when-writing-an-object-using
	La calidad mejora pero no en exceso, sigue grabando mucho ruido.	https://stackoverflow.com/questions/9389572/improve-audio-recording-quality-in-android
Librería para audios		https://exoplayer.dev/
Stream live audio to server	No hemos conseguido que funcione.	https://stackoverflow.com/questions/15349987/stream-live-android-audio-to-server
Proyecto desarrollado con Sockets de Android comunicándose con PHP.	Lo hemos testeado y no funciona.	https://stackoverflow.com/questions/29159354/php-socket-with-android

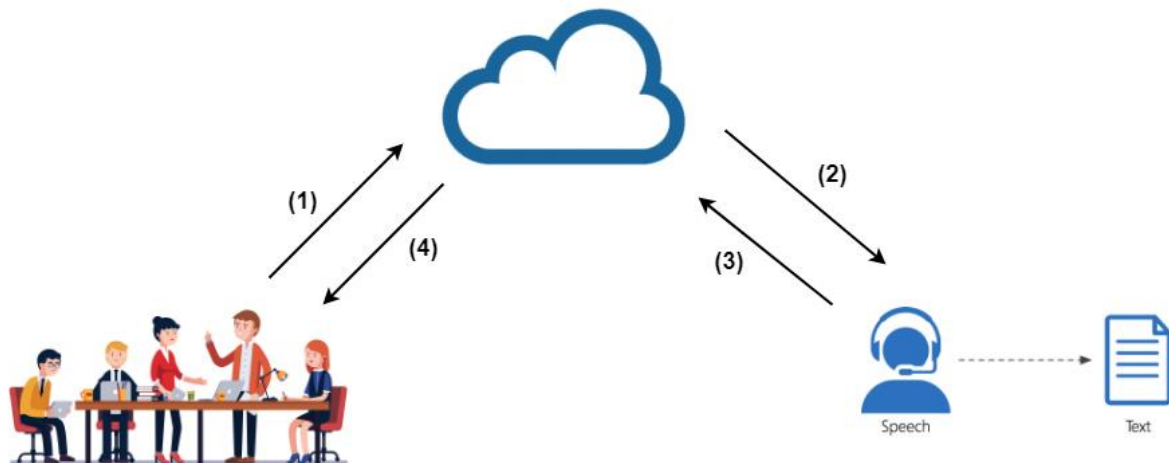
Ejemplo descargable.		https://stackoverflow.com/questions/15349987/stream-live-android-audio-to-server https://stackoverflow.com/questions/23497330/streaming-audio-from-android-app-to-php-server https://stackoverflow.com/questions/15349987/stream-live-android-audio-to-server
Web Speech API	Funciona muy bien en tiempo real, aunque sólo está disponible para una aplicación web que utilice JavaScript.	https://github.com/googlearchive/webplatform-samples
Aplicación de chat de audio en tiempo real mediante wifi	Sin tiempo para testearlo	https://github.com/Leewin0821/WiTalk
Tesis doctoral sobre separación de audios	Hemos leído gran parte del documento y la mayoría del pdf contiene información teórica y algunos algoritmos como BSS que podrían funcionar. Falta de tiempo como para investigarlo más a fondo.	https://arxiv.org/pdf/1808.10620.pdf

Hemos investigado las distintas librerías de código abierto de speech to text en tiempo real, algoritmos de detección de silencio y separación de audios (habiendo avanzado poco en estas últimas dos debido a la escasa información que existe).

Como conclusión, las librerías de transcripción de audio a texto en Android funcionan bastante mal (con mucho ruido y difícil de implementar en tiempo real). Hemos tenido muchos problemas al intentar hacer la aplicación en Android, y, al encontrar la **Web Speech API**, hemos decidido implementar esta solución puesto que es la que mejor funciona de todas las encontradas (con mucha diferencia). Es de Google y reconoce bien la mayoría de las palabras mencionadas.

Arquitectura

Analizando la solución inicial y habiendo realizado las investigaciones descritas anteriormente, se llegó a la conclusión de que la solución planteada era bastante compleja de desarrollar. La idea principal de recoger audios y enviarlos a un servidor para transcribirlos se mantiene, pero con algunos cambios como se ve en el siguiente esquema.



Donde:

- **(1):** El audio se recoge a través del dispositivo que utiliza cada asistente de la reunión, como puede ser un teléfono móvil, una tablet, un ordenador o un micrófono externo, y se envía a un servidor.
- **(2):** El servidor envía el audio en tiempo real a un módulo que lo transcribe a texto.
- **(3):** El módulo transcriptor devuelve los resultados al servidor.
- **(4):** El servidor envía el texto transcrito a todos los usuarios de la reunión.

La mayor diferencia que tiene este modelo con respecto al anterior, es que el reconocimiento de la huella de voz de cada persona y el procesamiento del audio para saber qué persona ha hablado, queda totalmente eliminado.

Con este modelo, se ahorra los costes de procesamiento de audio y de reconocimiento de huella digital mediante el uso de un dispositivo por asistente, como un teléfono o un ordenador, pudiendo cumplir el objetivo de que el chat de la reunión fuese en tiempo real. Al tener un dispositivo por asistente, es fácilmente reconocible qué persona ha estado hablando en la reunión mediante un sistema de usuarios.

Otras ventajas que se obtienen con este modelo frente a la solución inicial son las siguientes:

- No hay limitaciones físicas de los asistentes para poder participar de forma activa en la reunión, ya que la aplicación funcionará del mismo modo al estar desarrollada mediante el protocolo TCP/IP en lugar del Bluetooth.
- Ahorro de costes en fabricación de dispositivos externos como se plantearon inicialmente.
- Aplicación más escalable permitiendo que personas externas a la sala puedan participar también en la reunión.

Experimentos realizados

En este apartado se describen los experimentos que se han realizado durante el desarrollo de las aplicaciones.

Inicialmente se decidió desarrollar la aplicación en Android. Para ello se ha implementado diferentes librerías de audio en el ejemplo de una grabadora con el objetivo de grabar los audios, transcribirlo a texto visualizándolo por pantalla y analizar posteriormente la calidad de los mismos (tanto del audio como del texto).

Reconocimiento de voz

Tras analizar muchísimas librerías de código abierto de transcripción de voz a texto en Android, hemos visto que no funcionan tan bien como se esperaba. Se detecta bien el audio pero se guarda con mucho ruido lo que supone una muy mala transcripción a texto.

Calidad del audio

Esta prueba consistió en analizar la calidad del audio a través de una entrada de audio. Se utilizará el micrófono del teléfono móvil, un micrófono externo y dos personas (Persona 1 y Persona 2) para este experimento.

Para ambos dispositivos de recogida de audio, se realizarán tres pruebas teniendo el dispositivo de entrada de audio una de las ellas durante todo el experimento (para este ejemplo, la Persona 1):

- **Prueba 1:** La Persona 1 hablará a través de la entrada de audio.
- **Prueba 2:** La Persona 2 hablará sin tener la entrada de audio delante, estando la Persona 1 en silencio.
- **Prueba 3:** Las dos personas hablarán al mismo tiempo.

De cada una de las pruebas anteriores, se describe los resultados obtenidos que son los siguientes.

	Micrófono del teléfono	Micrófono externo
--	------------------------	-------------------

Prueba 1	La voz captada se escucha bien pero el nivel de voz es algo bajo.	Se recoge la voz perfectamente y con claridad.
Prueba 2	Estando a 0.5 metros, la voz se capta prácticamente igual que la Persona 1. A 1 metro la voz apenas recoge voz.	A 0.5 metros capta muy poca voz, casi inapreciable.
Prueba 3	La diferencia de volumen de voces era prácticamente la misma a 0.5 metros de distancia entre personas.	Recoge muy bien la voz de la Persona 1 y de la Persona 2 capta un audio muy bajo y apenas reconocible.

Chat en tiempo real

Debido a que los resultados obtenidos del experimento en la aplicación móvil no han cubierto las expectativas, se ha decidido implementar una aplicación web. El objetivo de esto era comprobar si las librerías de transcripción de voz a texto para aplicaciones web obtenían mejores resultados que antes.

Para ello, se ha decidido hacer algunas pruebas para comprobar el funcionamiento de la aplicación. El experimento consiste en un chat entre dos personas (Persona 1 y Persona 2) hablando a través del micrófono del portátil de cada uno.

Inicialmente, cada una de las personas se van a registrar en la aplicación, y a continuación, una de ellas creará la reunión y la otra se unirá a ella. A partir de este momento, cada uno de los usuarios activará su micrófono y empezará el chat entre ellos. Las pruebas realizadas han sido las siguientes:

- **Prueba 1:** El chat se realizó alternando el orden del hablante después de cada frase mencionada.
- **Prueba 2:** Las dos personas han hablado a través del mismo micrófono de un portátil, una en frente del micrófono y la otra separada de la primera aproximadamente medio metro.
- **Prueba 3:** Las dos personas hablarán continuamente sin importar si la otra ha dejado de hablar o no.

Estas mismas pruebas se han hecho incorporando un micrófono externo al portátil y éstos han sido los resultados:

	Micrófono del portátil	Micrófono externo
--	-------------------------------	--------------------------

Prueba 1	Resultados con una muy buena calidad de transcripción del audio identificando perfectamente a qué persona está hablando. El único fallo es cuando la Persona 2 está callada y la Persona 1 habla con un tono muy alto, mostrándose por pantalla el mensaje duplicado.	Resultados con una muy buena calidad de transcripción del audio identificando perfectamente a qué persona está hablando. Mejora sustancialmente el uso del micrófono externo.
Prueba 2	Resultados muy buenos, sólo en el caso de que la Persona 1 hable más alto que la Persona 2. En caso contrario, se mostraría una identidad errónea del mensaje de la pantalla.	En el caso de que la Persona 1 sea el que se ponga el micrófono externo y la Persona 2 no lo tenga, mejora mucho los resultados.
Prueba 3	Si las dos personas hablan con un tono de voz normal se muestran por pantalla los mensajes esperados. En el caso de que la Persona 1 hable con un tono de voz muy fuerte, sus mensajes se mostrarían perfectamente por pantalla mientras que los de la Persona 2 serían erróneos.	Mismo resultado que la prueba 1, con la diferencia de que ahora el factor del tono de voz de una persona puede afectar a la integridad de los mensajes.

Uso de librería Web Speech API

De todas las librerías de transcripción de voz a texto investigadas, la que ofrecía un mejor rendimiento y en tiempo real era Web Speech API.

Esta API se probó en un ordenador con el micrófono interno y con un micrófono externo y los resultados eran bastante buenos.

Sin embargo, esta librería sólo puede utilizarse en navegadores Chrome, siendo la versión mínima la 25.

Las pruebas realizadas con teléfonos móviles que se conectan a la aplicación mediante el navegador Chrome, utilizando tanto el micrófono del teléfono móvil y uno externo, no han sido muy buenas. El audio que se recoge a través del teléfono se recoge varias veces, es decir, si el interlocutor dice “Hola buenos días”, el micrófono del teléfono recogerá el mismo mensaje varias veces seguidas, y se mostrará en el mensaje que se envía al chat de la reunión (véase la Figura 1). Este problema parece ser que afecta únicamente a los teléfonos móviles, por lo que el uso de esta API debería ser utilizada en un ordenador.

(10:54) *Emilio dice:*

Esto es un fallo en los teléfonos Androidesto es un fallo en los teléfonos Android

Figura 1. Error de Web Speech API en teléfonos Android.

Desarrollo

Tras las investigaciones realizadas y comprobar que la tecnología TCP/IP ofrecía más ventajas frente a la tecnología Bluetooth, se ha creado una aplicación web que soporta esta tecnología.

La aplicación web se ha desarrollado con la arquitectura MVC (Modelo Vista Controlador) desde cero, ya que debido al tiempo que teníamos, no podíamos retrasarnos mucho tiempo en aprender a utilizar un framework de PHP como Symfony o Laravel. Con esta filosofía, se ha querido mantener la arquitectura lo más limpia y legible posible para que con pocas líneas de código se realicen las funciones principales y además, ayudar al propio mantenimiento del software.

La aplicación web contiene tres componentes principales, que son:

- **Sistema de usuarios:** Mediante un login y registro de usuarios, podrán acceder a la aplicación y cerrar sesión. Este sistema es el utilizado para identificar a cada usuario cuando recoja la voz y ponga la transcripción a texto en el chat en tiempo real.
- **Transcripción de audio:** Se ha utilizado un código open-source llamado Web Speech API para la recogida de audio y la transcripción a texto. Está hecho en JavaScript, su funcionamiento es enviar a través de streaming la voz que recoge a un servidor de Google y devuelve el texto asociado a dicho audio en tiempo real. La calidad del reconocimiento se acerca bastante al 100%, por lo que en un futuro habrá que cambiar este módulo por otro mejor.
- **Chat en tiempo real:** Para realizar un chat asíncrono, se ha utilizado la tecnología Ajax y JQuery. Los mensajes captados por la transcripción de audio a texto, se envían al servidor para que se guarden en la base de datos, junto con el nombre de usuario que ha hablado. El funcionamiento es muy simple, se ha realizado un bucle infinito que se ejecuta cada segundo que actualiza los mensajes para poder ver así los mensajes que se envían en tiempo real.

En cuanto a las vistas, se ha utilizado un motor de plantillas PHP llamado Twig (<https://twig.symfony.com/>), que reutiliza el código de las vistas y resulta mucho más cómodo de mantener.

Como se utiliza la librería Twig, se ha añadido un controlador de dependencias llamado composer, que al instalar la aplicación, habrá que ejecutar también la instalación de estas dependencias para que funcione.

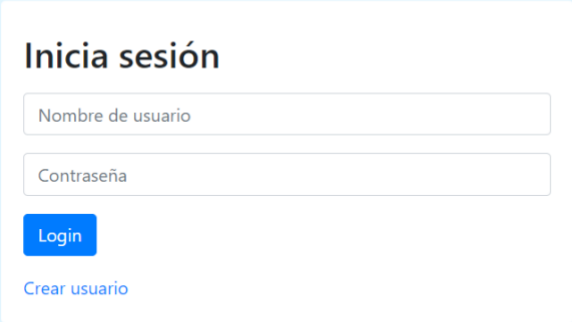
Para evitar que se mezcle el audio entre los asistentes de la reunión, lo mejor es utilizar micrófonos externos. Las pruebas realizadas como se ha comentado anteriormente, han concluido que además de filtrar mejor las voces entre los interlocutores, aumenta la calidad del audio recogido, por lo que será más sencillo para el módulo poder transcribir la voz a texto.

El código fuente se puede descargar de <https://github.com/osoc-es/WhoTalks>

Interfaz

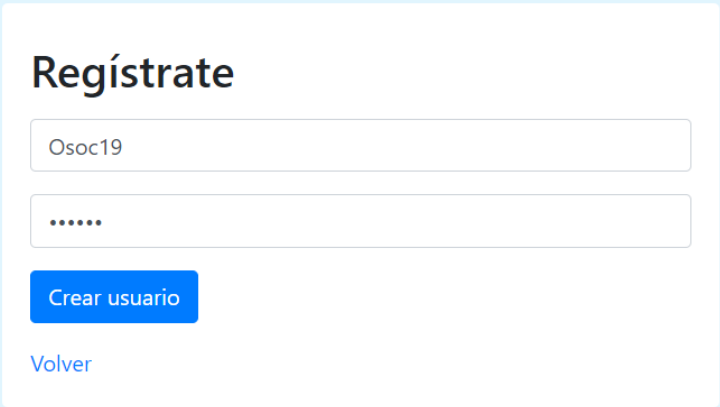
A continuación, se muestra el diseño inicial de las interfaces de la aplicación que son las siguientes.

En las Figuras 2 y 3, se puede ver el inicio de la aplicación, donde se ha desarrollado un sistema de usuarios para tener más seguridad en las partes de la aplicación y que además, nos sirve para identificar a cada usuario en las reuniones que se realicen.



The image shows a login interface titled "Inicia sesión". It features two input fields: "Nombre de usuario" and "Contraseña". Below the fields is a blue "Login" button. At the bottom, there is a link that says "Crear usuario". The entire form is set against a light blue background.

Figura 2. Login de usuarios en WhoTalks.



The image shows a registration interface titled "Regístrate". It features two input fields: the first contains the text "Osoc19", and the second is a password field represented by dots. Below the fields is a blue "Crear usuario" button. At the bottom, there is a link that says "Volver". The entire form is set against a light blue background.

Figura 3. Registro de usuarios en WhoTalks.

En la Figura 4 se muestra el inicio de la aplicación, donde se pueden crear reuniones rápidamente para iniciar un chat.



Figura 4. Pantalla de inicio.

En las Figuras 5 y 6 se muestra el chat de la reunión, donde para poder participar en la reunión, se debe activar el micrófono pulsando el botón 'Micrófono' y aceptar los permisos de uso del micrófono.

A la izquierda, se muestra la lista de usuarios que están participando en la reunión y se irá actualizando cuando se conecte un nuevo usuario a la reunión. Debajo, se encuentra un botón para salir de la reunión y también un recuadro oscuro que contendrá el resultado del reconocimiento de la voz de la persona.

A la derecha se muestra el listado de los mensajes de la reunión, en el que una vez activado el micrófono, se irá recogiendo la voz para transcribirla a texto y se enviará al chat en tiempo real.

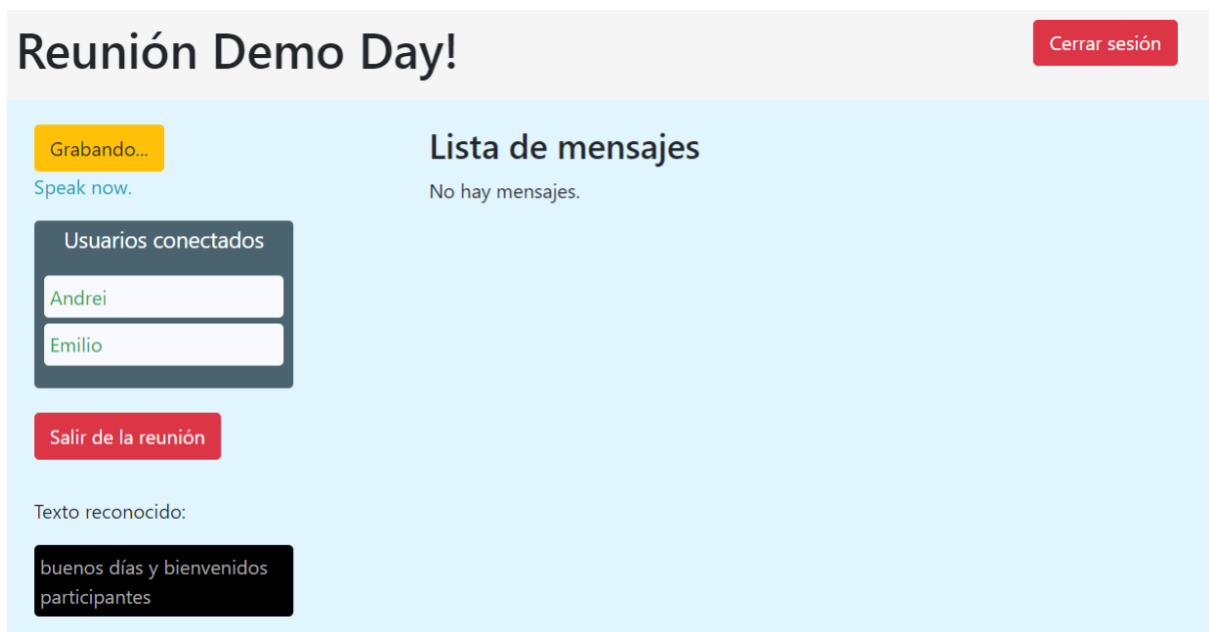


Figura 5. Reconocimiento de voz.

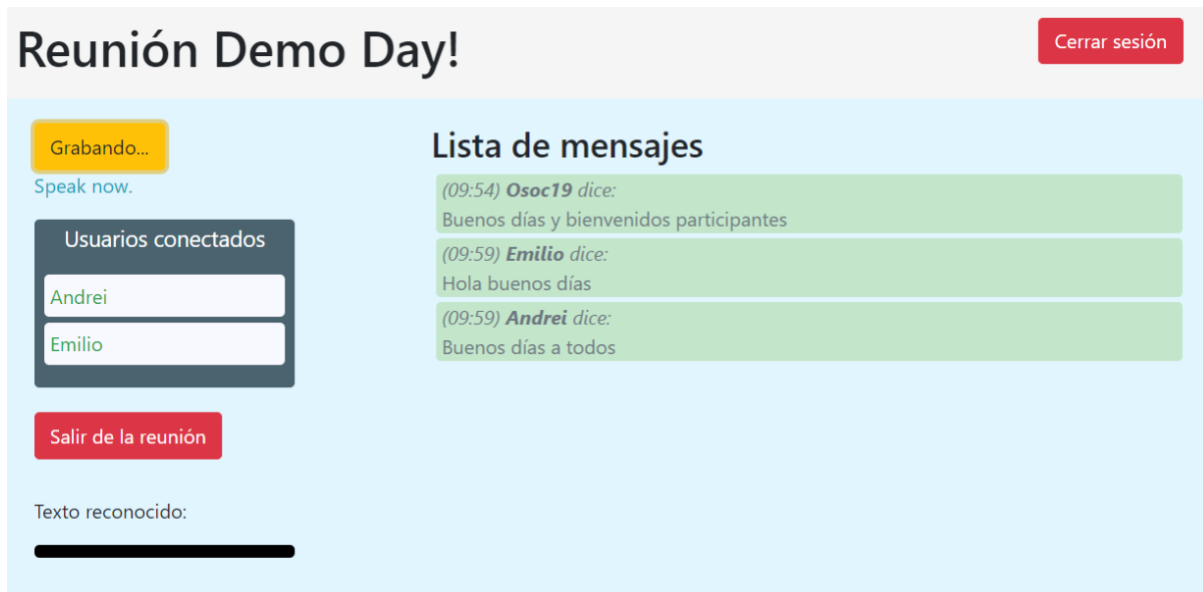


Figura 6. Mensajes de chat con varios usuarios.

Dificultades encontradas

Investigación

Este proyecto es muy innovador y presenta un problema real que se produce actualmente, por lo que nos ha sido difícil encontrar artículos de calidad gratuitos que nos pueda ayudar para desarrollar el proyecto. La mayoría de artículos eran poco específicos siendo muy teóricos especialmente en la parte de *Separación de Audio de diferentes personas*. Debido a esto no hemos podido implementar ningún algoritmo capaz de separar las distintas voces presentes en un audio.

Desarrollo

En esta parte hemos tenido problemas, sobre todo, en la parte de Android. A la hora de probar las librerías de voz investigadas se producían un montón de errores (tanto de compilación como de ejecución) teniendo que descartar muchas de ellas sin estar convencidos del todo de si realmente eran útiles o no.

Conclusiones

Los objetivos propuestos inicialmente para este proyecto se han cumplido satisfactoriamente por las siguientes razones:

La aplicación debe desarrollarse en tiempo real, y para ello se ha cambiado la tecnología Bluetooth por TCP/IP para cumplir este cometido. Debido a este cambio, hemos aumentado su funcionalidad, ya que ahora las reuniones no estarán limitadas físicamente y cada asistente puede encontrarse en cualquier lugar del mundo para poder participar en las reuniones.

La separación de audios mediante huella de voz era bastante complejo de realizar sin que esto afectase al rendimiento computacional. En la solución propuesta, no es necesario separar el audio para identificar a un usuario, sino que se utilizará un dispositivo por cada asistente a la reunión, como puede ser un ordenador, tablet o móvil, que identifica a un usuario y con ello, se ahorra en costes computacionales.

La transcripción de audio a texto se ha realizado mediante una api de Google llamada Web Speech API de tipo open-source que funciona bastante bien a la hora de reconocer palabras y sobre todo y lo más importante, que el audio lo transcribe de forma instantánea.

Y por último, la funcionalidad implementada en el chat mediante la tecnología Ajax, hace que los mensajes tanto los recibidos por otros usuarios como los enviados sean en tiempo real.

En cuanto a los trabajos futuros, se puede mejorar este proyecto con los siguientes cambios propuestos:

- La interfaz puede mejorarse debido a que la propuesta es un diseño inicial, además, podrían incorporarse nuevas funcionalidades como personalizar los colores y tamaños de letra que aparecen en el chat para que sea más cómodo al usuario. También, podrían señalarse con otro color las palabras clave que se escriben en el chat para poder realizar una lectura rápida del chat de la reunión.
- Como se almacena todos los mensajes que hay en una reunión, se podría implementar una funcionalidad en la aplicación que genere el acta mediante un documento word o pdf para poder analizarlo posteriormente.
- El módulo transcriptor que se ha implementado todavía no ofrece una garantía del 100% en el reconocimiento de voz, por lo que en un futuro podría cambiarse por otro módulo transcriptor que mejore las prestaciones actuales.
- Relacionado con lo anterior, se podrían investigar nuevas técnicas que mejoren la calidad de audio para que los futuros módulos transcriptores ofrezcan resultados más precisos.

