Especificaciones técnicas

GENERACIÓN DE
VOCES EN CATALÁN MEDIANTE I.A.

1 Introducción

Como parte de las actuaciones de la dirección de Estrategia Tecnológica en RTVE, en el año 2023 se llevaron a cabo dos pruebas de concepto relacionadas con la información electoral en poblaciones pequeñas, de menos de 1.000 habitantes, para las elecciones municipales y generales que, basándose en tecnologías de Inteligencia Artificial (IA) y a partir de los datos suministrados por las fuentes de información oficiales, los interpretaba y presentaba una noticia creada automáticamente.

Para mejorar el alcance en la difusión y acceso a la información, las PoC incluían en cada noticia la generación de un texto para leer que podría reproducirse con voces sintéticas únicas desarrolladas partiendo de voces humanas correspondientes a profesionales de la comunicación.

Partiendo de los conocimientos adquiridos en estas PoC se está trabajando en un nuevo proyecto de investigación aplicada para generar noticias de otras temáticas que incluyan texto y voz en dos idiomas.

2 Lote único

Mediante este expediente se contratará el desarrollo de un software de generación de **voces sintéticas únicas en catalán** (masculina y femenina) creadas partiendo de voces humanas y un servicio que, utilizando estas voces, genere piezas de audio que acompañen a las noticias sobre información meteorológica, creadas mediante IA para localidades de menos de 10.000 habitantes de la provincia de Lérida.

Para ello, se trabajará en el software de generación de voces en catalán con acento central y acento del noroeste en los primeros meses del proyecto. Posteriormente, durante tres meses, se generarán diariamente audios para incorporar a noticias textuales de carácter meteorológico.

El adjudicatario de este expediente deberá garantizar la conectividad con la empresa responsable de la generación de textos que le será comunicada por RTVE una vez adjudicado el proyecto, asegurando que el sistema ofertado tenga las prestaciones suficientes para que no se colapse el flujo de datos ni se produzcan retrasos en la entrega de la noticia.

Para cada población se creará una versión en castellano y otra en catalán. Además de las voces, los audios incorporarán efectos sonoros relacionados con los fenómenos meteorológicos (truenos, lluvia, viento) así como una sintonía que serán proporcionados por RTVE.

La empresa encargada de la generación de textos creará uno específico, con las características adecuadas de duración y formato del contenido, para su transcripción a voz, para cada una de las noticias, y en cada uno de los idiomas. Con estos textos, el adjudicatario de este expediente generará los audios correspondientes, así como clips en los que las voces aparezcan mezcladas con las sintonías y efectos sonoros. Estos clips deberán estar disponibles y publicados en la web www.rtveia.es en el mismo momento en el que las noticias sean accesibles para el usuario final, en un tiempo no superior a 15 minutos desde la generación de los textos.

Será condición imprescindible para aceptar la oferta que **los oferentes dispongan en el momento de presentar la oferta de lo siguiente:**

- ❖ Al menos dos voces sintéticas únicas, una masculina y otra femenina, que conviertan el texto de noticias de corta duración a voz (Text to Speech – TTS) en español con acento de España con acento del centro peninsular. Los modelos de voz deberán cumplir las siguientes características:
  - ➢ Se habrán generado partiendo de **voces humanas** clonadas **de periodistas**, sin embargo, no deben identificar a ninguno concreto.
  - ➢ Clonadas en el tono de voz y en lo referente a la manera de leer los textos, entonación, cadencia, énfasis, manera de leer los números, etc. Se trata de evitar voces fuente robóticas, poco naturales o diseñadas para otros menesteres.
  - ➢ Incluirán y podrán controlar atributos relacionados con las propiedades características de la voz en términos de prosodia tales como: ritmo, cadencia, pausas de respiración, velocidad, tono y timbre y estar generados con las técnicas más actuales del estado del arte de la síntesis de voz.
  - ➢ La naturalidad de la voz debe tener un MOS (Mean Opinion Score) superior a un 4.

- ❖ Capacidad de generar con estas voces, las piezas de audio que incorporen las sintonías asociadas y los efectos sonoros en los planos de tiempo correctos.
- ❖ Generar los audios en tiempo real.

RTVE podrá solicitar a todos los oferentes, si lo considera oportuno, una demostración de que cumplen las condiciones anteriormente mencionadas.

A continuación, se describen las características técnicas y funcionales con las que deberá contar el servicio ofertado.

- o Reentrenamiento de los modelos a partir de grabaciones de voces consensuadas con RTVE de forma que se disponga de voces tanto en español como en catalán. En el caso del catalán una de las voces tendrá acento central y la otra, acento del noroeste. RTVE decidirá una vez adjudicado el expediente la voz, masculina o femenina, que considere más adecuada para cada acento.
- o Personalización del servicio a partir de la creación de herramientas NLP y diccionarios fonéticos en castellano y catalán para adecuar la síntesis al vocabulario de información meteorológica. Ajuste del sistema y puesta en funcionamiento:
  - ▪ Infraestructura de pruebas. Estará disponible desde la firma del contrato. Los audios generados en las diferentes pruebas descritas anteriormente se quedarán almacenados y disponibles durante la vigencia del contrato de forma que se puedan realizar cuantas demostraciones del funcionamiento del sistema se consideren necesarias por parte de RTVE.
  - ▪ Infraestructura para la prueba de concepto. Esta infraestructura, con capacidad suficiente para atender las demandas durante los meses de servicio, deberá almacenar los audios generados hasta una semana después de la finalización del servicio.
- o Disponibilidad del servicio en tiempo real y 24/7:
  - ▪ Infraestructura de pruebas: durante la fase inicial hasta el inicio del servicio.
  - ▪ Infraestructura de la PoC: una semana antes del comienzo del servicio hasta una semana después de finalizado el mismo.
- o Los audios generados se incorporarán a la página Web seleccionada por RTVE, donde se podrán escuchar en una sola voz, con una probabilidad del 50% para voz masculina y 50% para la femenina. Para el caso del castellano, las dos voces tendrán el mismo acento, para el catalán, los acentos serán los descritos anteriormente.
- o Las noticias generadas, serán de duración corta (20 segundos aproximadamente), e incluirán sintonías y efectos sonoros proporcionados por RTVE.

- Las noticias con información meteorológica se publicarán en la página Web , con una periodicidad diaria, para ello, el adjudicatario convertirá a voz los textos recibidos de la empresa que genera los mismos, creará los clips correspondientes a cada municipio, en español y en catalán y enviará a dicha empresa una URL con el audio en mp3 por cada clip generado.
- Se realizarán cuantos reentrenamientos de los modelos sean necesarios hasta conseguir la correcta pronunciación en catalán.
- Los audios en castellano deberán mantener la pronunciación de los municipios en catalán.
- Las voces generadas para este proyecto serán propiedad de RTVE y no podrán ser utilizadas para ninguna otra prueba, proyecto o actividad ajena a RTVE.

2.1 Características del servicio

A partir de la fecha de la firma del contrato, el adjudicatario dispondrá de un periodo de tiempo para el desarrollo del software para la generación de las dos voces en catalán según las características descritas anteriormente que finalizará el 31 de agosto de 2024.

Además, durante este periodo, el adjudicatario participará en las pruebas conjuntas con la empresa generadora de textos, para lo cual, deberá adaptarse a la planificación establecida por el equipo de trabajo RTVE, asistiendo a las reuniones de coordinación junto con la empresa responsable de la generación de textos.

Estas pruebas le servirán de referencia para la adecuación y mejora de su herramienta permitiéndole llegar al inicio del servicio con la calidad adecuada.

El servicio comenzará el 1 de septiembre y se prolongará por tres meses durante los cuales se generarán diariamente los audios que acompañarán a las noticias textuales sobre información meteorológica para las poblaciones de Lérida de menos de 10.000 habitantes.

Aunque el servicio finaliza en noviembre, los archivos de audio del TTS de las voces deberán almacenarse hasta el 31 de diciembre del 2024, entregándose los mismos a RTVE a la finalización del contrato.

3 Coordinación con CRTVE

La empresa adjudicataria nombrará un responsable del servicio dentro de su equipo, que será el coordinador. Este coordinador será el interlocutor único con el personal de CRTVE para la coordinación y priorización de tareas. Así mismo, desde CRTVE también habrá otro coordinador que será su único contacto para cualquier duda o problema que pueda surgir. Se establecerán reuniones ordinarias semanales para el seguimiento de las tareas. Será labor de los coordinadores la convocatoria de reuniones extraordinarias si se considerara necesario.

4 Formato de la oferta

La oferta técnica presentada tendrá la estructura que se indica a continuación:

- Organización del trabajo y sistema de aseguramiento de la calidad del servicio.
- Cumplimiento y mejoras de las especificaciones técnicas requeridas.

- Relación de recursos humanos, aportando documentación que acredite la experiencia solicitada en el Pliego de Condiciones Generales.
- Cumplimiento y mejoras de las especificaciones técnicas requeridas. En este sentido, se incluirá en la Oferta Técnica todos los documentos que acrediten el cumplimiento de los criterios técnicos de valoración de las ofertas descritos en el Pliego de Condiciones Generales.
- Planificación de tiempos, lo más detallada posible de los plazos de entrenamiento y aprendizaje del sistema, especialmente en catalán, pruebas y puesta en servicio.