Especificaciones técnicas

GENERACIÓN DE VOCES Y ASISTENTE VIRTUAL PARA LA COBERTURA INFORMATIVA DE LAS PROXIMAS ELECIONES GENERALES

1 Introducción

Como parte de las actuaciones en el ámbito de la IA aplicada a la producción audiovisual en RTVE, y amparado por el expediente S-01392-2023 "I.A. Generación automática de textos", se está llevando a cabo un proyecto de investigación aplicada de un sistema que, basándose en tecnologías de Inteligencia Artificial (IA) y a partir de los datos suministrados por las fuentes de información oficiales que se determinen, los interprete y presente una noticia creada automáticamente, con texto en lenguaje natural y relativa a información electoral en poblaciones pequeñas, inicialmente de menos de 1.000 habitantes. Este proceso se hace sin intervención humana alguna, aunque siempre supervisado por profesionales para certificar y verificar su calidad informativa, editorial y estética.

En la fase actual del proyecto y con el objetivo de mejorar su alcance en el ámbito de la difusión y acceso a la información, se han realizado pruebas piloto y una producción completa en las elecciones municipales de 2023:

- Para el desarrollo de voces sintéticas capaces de reproducir el contenido de las noticias previamente elaboradas,
- Para la creación automática con dichas voces de piezas completas, en las que aparezcan mezcladas las sintonías de los programas con dichas voces.

Aportando así un valor añadido a la herramienta de creación automática de noticias.

Tras el éxito de los resultados obtenidos durante las pruebas piloto y la cobertura informativa de las elecciones municipales en poblaciones de menos de 1000 habitantes, la necesidad de seguir avanzando en materia de calidad y accesibilidad, se desea implantar esta funcionalidad para la cobertura de las elecciones generales en poblaciones de menos de 1000 habitantes, con una correcta implementación en la personalización de las voces, la incorporación de un asistente virtual capaz de leer las noticias e interactuar con el lector, así como la capacidad de pronunciar en otros dialectos y/o lenguas co-oficiales los municipios de interés.

2 Objeto del contrato

Mediante este expediente se contratará un servicio que utilizando **voces sintéticas únicas** (masculina y femenina) creadas partiendo de voces humanas genere piezas en las que además de la voz se incorpore la sintonía con la **pronunciación** de los distintos municipios, y de los nombres de algunos de los candidatos, en otros **dialectos y/o lenguas co-oficiales,** para su incorporación en una página Web. El servicio incluirá además un sistema de **experiencia conversacional**, para el que se contará con una **voz para asistente virtual**, generada igualmente partiendo de voces humanas.

Los resultados de estos desarrollos serán utilizados en la prueba de concepto que RTVE llevará a cabo para la generación de noticias de las elecciones generales que se celebrarán el próximo 23 de julio de 2023. Para ello, la empresa adjudicataria de este pliego deberá garantizar la conectividad con la empresa Narrativa, responsable de la generación de textos con las prestaciones suficientes para que no se colapse el flujo de datos ni se produzcan retrasos en la entrega del audio de la noticia. Igualmente, será responsabilidad del adjudicatario de este expediente los desarrollos necesarios para extraer los datos utilizados en la elaboración de las respuestas del asistente virtual de los ficheros proporcionados con los datos oficiales durante la PoC.

3 Lote único

El adjudicatario de este expediente garantizará que el sistema ofertado pueda dar servicio de forma holgada al alcance de la PoC antes mencionada, en la que se generarán 3 noticias en el día electoral para cada una de las aproximadamente 5.000 poblaciones españolas con menos de 1.000 habitantes. Previsiblemente una de las noticias se creará a lo largo de la tarde y las otras dos una vez cerrados los colegios electorales, con los resultados parcial y final de los comicios, no obstante, esta planificación podrá cambiarse por parte de los responsables de RTVE si las circunstancias así lo requieren.

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. Con estos textos, el adjudicatario de este lote generará los audios que los acompañan, así como clips en los que las voces aparecen mezcladas con las sintonías, que deberán estar disponibles 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, tanto el audio para publicación en web como el servicio del asistente virtual.

Para garantizar el buen funcionamiento conjunto de los sistemas que intervendrán en la PoC, hasta la fecha prevista para las elecciones se realizarán al menos las siguientes pruebas previas en las que el adjudicatario tendrá que participar y demostrar que cumple con lo solicitado en este expediente:

Pruebas de flujo de trabajo: se realizarán al menos dos hasta la fecha de las elecciones, con un alcance aproximado de 100 municipios

Pruebas de carga: se harán para todos los municipios contemplados en el alcance de la PoC y en ellas se reproducirá el escenario de la tarde-noche electoral, utilizando para ello resultados reales de comicios anteriores. Está previsto realizar un mínimo de dos pruebas de este tipo.

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

a) Al menos de dos voces sintéticas únicas, una masculina y otra femenina, que conviertan el texto de noticias de corta duración a voz (Text to Speech – TTS). 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 no solo en cuanto a tono de voz, sino también en lo referente a manera de leer los textos, entonación, cadencia, énfasis, manera de leer los números. Se trata de evitar voces fuente que sean robóticas, poco naturales o diseñadas para otros menesteres.
   - Deben incluir y poder 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.
   - Debe disponer así mismo de tecnología para generar nos solo las voces, sino también incorporar las sintonías asociadas y en los planos de tiempo correctos.
   - La naturalidad de la voz debe tener un MOS (Mean Opinion Score) superior a un 4.
   - Deben generar audios en tiempo real.

b) Al menos dos productos completos de asistente virtual, tanto las voces como la estructura completa de navegación, y al menos uno de ellos será de información sobre elecciones.

c) Capacidad y experiencia probada en la creación de piezas en las que además de la voz se incorpore la sintonía y la **pronunciación** de los distintos municipios objeto de esta PoC y de los nombres de algunos de los candidatos, en otros **dialectos y/o lenguas co-oficiales**.

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.

3.1 Síntesis de voz: Web

Durante el presente contrato se llevarán a cabo las siguientes tareas:

- Reentrenamiento de los modelos a partir de grabaciones de voces consensuadas con RTVE.
- Personalización del servicio a partir de la creación de herramientas NLP y diccionarios fonéticos en castellano para adecuar la síntesis al vocabulario electoral (municipios, partidos políticos, siglas y acrónimos...).
- 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 la capacidad suficiente para atender las demandas del día de las elecciones deberá estar disponible en el momento de la adjudicación y almacenar los audios generados hasta una semana después de la celebración de las elecciones (aproximadamente 15.000 noticias).
- Disponibilidad del servicio en tiempo real y 24/7:
  - Infraestructura de pruebas: durante la fase inicial hasta la noche electoral.
  - Infraestructura de la PoC: una semana antes del día de las elecciones y una semana después de éstas.
- Almacenamiento de la última versión de al menos 3 audios de noticias por municipio en producción (el día de las elecciones del municipio) y durante al menos 7 días tras las elecciones. Las voces generadas se incorporarán a la página Web de pruebas proporcionada por RTVE, donde se podrán escuchar en una sola voz o a dos voces, según será definido por la dirección del proyecto de RTVE.
- Las noticias generadas, serán de duración corta (20 segundos aproximadamente), e incluirán sintonías proporcionadas por RTVE.
- La tarde-noche electoral se publicarán los resultados en la página Web designada por parte de RTVE.
- El adjudicatario convertirá a voz los textos recibidos de la empresa que genera los mismos creará el clip correspondiente a cada municipio y enviará a dicha empresa una URL con el audio en mp3 por cada clip generado.

3.2 Síntesis de voz: asistente virtual

El adjudicatario proporcionará un asistente virtual para una experiencia accesible más allá de la Web que cumpla los siguientes requisitos:

- implementación del asistente: capacidad de generar audios sobre las noticias de un municipio que esté incluido en la prueba a partir de la petición de un usuario. El usuario escuchará los audios con una voz de las solicitadas en el punto 3.1.

- Posibilidad de escuchar la noticia por partes en función de los deseos del usuario mediante relación conversacional con el asistente virtual.
- Creación de una Skill para el asistente conversacional Alexa.
- Envío a certificación y publicación de la experiencia en la tienda de Amazon al menos una semana antes de la noche electoral.
- Revisión y mantenimiento del NLU en producción para optimizar la prosodia asociada a los nombres de los municipios.

3.3 Voz para asistente virtual

Aportación de una tercera voz sintética, adicional a las dos recogidas en los puntos anteriores, exclusiva para la experiencia en el asistente virtual con las mismas características que en el punto 3.1. Además, la voz será utilizada puntualmente en la experiencia de voz del asistente virtual, como sustituta a la voz de Alexa.

3.4 Pronunciación de los municipios en otras lenguas dialectos y/o lenguas co-oficiales

Mejora de los modelos de síntesis de voz generados mediante:

- Re-entrenamiento de los modelos con nuevas grabaciones
- Mejora del diccionario fonético para la pronunciación de los municipios adaptada a otros dialectos y/o lenguas co-oficiales.

3.5 Características del servicio

A la firma del contrato el adjudicatario dispondrá de las voces sintéticas completamente adaptadas a las necesidades planteadas en el presente pliego hasta su integración en el servicio que se prestará durante la noche electoral. La duración del contrato será hasta 7 días después de la noche electoral de las elecciones generales de 2023 (prevista para el 23 julio de 2023).
Los archivos de audio del TTS de las voces deberán almacenarse, entregándose los mismos a RTVE a la finalización del contrato.
El adjudicatario intervendrá en las pruebas de campo contempladas hasta dicha fecha y asociadas al proyecto del expediente S-01392-2023, anteriormente mencionado, y deberá adaptarse a la planificación establecida por el equipo de trabajo RTVE

4 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.

5 Formato de la oferta

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

- Compromiso de calidad del servicio.
- Organización del trabajo y sistema de aseguramiento de la calidad del servicio.
- Cumplimiento y mejoras de las especificaciones técnicas requeridas.
- Planificación de tiempos, lo más detallada posible de los plazos de entrenamiento y aprendizaje del sistema, pruebas y puesta en servicio.