Entrenamiento de modelos para la generación de "deep-fakes" de voz en idioma castellano.

Daniel Doña Álvarez

Grado en Ingeniería Informática

Área de Inteligencia Artificial

Consultor: Ferran Diego Andilla

Profesor responsable: Carles Ventura Royo

Contenidos

1 Introducción **3**

2 Descripción del TFG **3**

3 Objetivos generales y específicos **4**

4 Planificación con hitos y temporalización **4**

4.1 Organización temporal 4

4.2 Hitos de investigación 5

4.3 Hitos prácticos 5

4.4 Riesgos contemplados 6

5 Contenido de las PECs y de la entrega final **6**

5.1 PEC 2 6

5.2 PEC 3 7

5.3 PEC 4 7

5.4 PEC 5a 7

6 Estructura de la memoria del TFG **7**

1 Introducción

En la entrega anterior se abordó ya una aproximación al ámbito de investigación en el cual se movería el trabajo realizar. Se aproximó también en cierto grado el problema que aspiraba a abordar y arrojar cierta luz.

En esta entrega por tanto se profundizará en ofrecer ya una descripción más madura del trabajo, así como una organización de los objetivos y su expresión en tareas temporalizadas.

2 Descripción del TFG

Entendemos por síntesis del habla a todas aquellas formas artificiales que reproduzcan la fonética de la comunicación humana. Los primeros intentos de lograr esta síntesis datan de siglos atrás, cuando se consiguió la articulación de fonemas en base a reproducciones anatómicas del tracto vocal y el uso de fuelles.

Aquellos modelos mecánicos quedan muy lejos del paradigma actual, para acercarnos a la base de la que pueden partir nuestro trabajo tenemos que remontarnos a los años 30 del siglo pasado, cuando en los laboratorios Bell se desarrollaba por Homer Dudley lo que hoy llamamos vocoder. Así se conseguía por fin codificar la voz humana en señales eléctricas e igualmente usar esas señales eléctricas para reproducir en cierto grado la voz original.

Hoy en día la técnica ha avanzado bastante y los sistemas pneumáticos o la modificación laboriosa de señales analógicas lo hemos reemplazado con computadores y modelos entrenados. Las últimas dos décadas y en especial los últimos años nos han dejado avances bastante prometedores en este sentido tanto en la calidad de los resultados como en la posibilidad abierta de entrenamiento y mejora de dichos modelos.

El marco de trabajo es por tanto la síntesis del habla en base a modelos de aprendizaje computacional, pero el tema concreto es algo ligeramente más acotado. Las aplicaciones prácticas de una síntesis del habla realista son muchas, desde apoyo a personas con diversidad funcional hasta los altavoces inteligentes que se han popularizado entre el público general. Pero no todas las aplicaciones son tan constructivas ni inocentes como cabría esperar.

En la medida en que las capacidades de cómputo crecen y podemos diseñar y entrenar mejores modelos, cada vez más se difumina la diferencia entre aquello que es creado genuinamente por una persona y aquello que es la inferencia de un modelo entrenado. En la posibilidad real de presentar lo segundo como lo primero es donde aparecen los llamados "deep-fakes".

Estas falsificaciones realistas pueden ser de muchos tipos, pero su uso se ha generalizado en aquellas centradas en contenido audiovisual. Esto no siempre se ha realizado con fines cuestion-

ables, también han encontrado un hueco en el espacio de la creación artística, especialmente en el cine. Pero sí hay que señalar que ha existido y existe su uso como vehículo de información falsa.

Así pues el tema de investigación se centra en estas falsificaciones del habla a partir del estado del arte de la síntesis de voz y más específicamente en el idioma castellano. Adicionalmente desean abordar no solo los modelos existentes para su generación sino también sus antagonistas: las herramientas existentes para su detección.

3 Objetivos generales y específicos

Para la organización del tema a abordar se ha decidido seguir el siguiente esquema de objetivos:

- Estudio de los modelos existentes para la producción de voz a partir de texto
  - Estudio de los modelos TTS entrenados en general
  - Estudio de otros modelos de síntesis de habla
  - Estudio específico de los modelos orientados a generar deep-fakes
- Estudio de las herramientas existentes para la producción de deep-fakes
  - Evaluación de las herramientas existentes
  - Producción de un deep-fake usando mi voz
- Estudio de la teoría y formas de detección de deep-fakes

4 Planificación con hitos y temporalización

4.1 Organización temporal

Para la organización temporal se han tenido en cuenta las planificaciones orientativas de las entregas siguientes así como el retraso acumulado hasta el punto actual con esta entrega.

Se considera un período efectivo de trabajo de dos meses, siendo estos abril y mayo. Dejando el mes de junio para la redacción final de la memoria así como la preparación de la presentación del trabajo.

Por otra parte, se han identificado una serie de dependencias entre las tareas derivadas de los objetivos marcados. Estas dependencias no se considera que impidan la paralelización de

las tareas pero sí marcan el grado de profundización de unas respecto a las otras.

Se ha acordado la realización de reuniones regulares con el consultor, siendo la frecuencia de dos semanas salvo necesidades urgentes sobrevenidas.

El tiempo de dedicación personal semanal se ha fijado en un mínimo de 20 horas semanales, distribuidas en 10 horas los días laborables y 10 horas los fines de semana.

4.2 Hitos de investigación

Por lo complicado de medir la consecución de hitos en el ámbito teórico la mejor forma de cerrar estos hitos es la preparación de un borrador o esquema de los conocimientos adquiridos en ellos como base para la redacción

- Recopilación de una base esencial de la literatura existente como punto de partida para la investigación: 25 de marzo
- Conocimiento básico de los modelos existentes en TTS: 8 de abril
- Conocimiento de las características más relevantes de los modelos para la producción específica de deep-fakes: 15 de abril
- Conocimiento de las técnicas para la detección de deep-fakes: 15 de mayo

4.3 Hitos prácticos

- Preparación del entorno de trabajo para el entrenamiento: 27 de marzo
  - Configuración de CUDA, Torch, Tensorflow y otras piezas software.
  - Pruebas empleando hardware propio y valoración de usar Google Colab u otros servicios similares.
- Entrenamiento de modelos existentes con datasets existentes: 15 de abril
  - Tacotron2
  - WaveRNN
  - WaveGlow
- Generación de un dataset propio de voz para el entrenamiento de modelos: 15 de abril
- Entrenamiento de modelos existentes con el dataset generado: 30 de abril

  - Tacotron2
  - WaveRNN
  - WaveGlow

- Posible mejora y ajuste de modelos estudiados al dataset generado: 15 de mayo
- Prueba de herramientas de detección de deep-fakes: 15 de mayo
- Despliegue de una interfaz de prueba de los modelos generados: 30 de mayo.

4.4 Riesgos contemplados

En la planificación anterior existen riesgos notables, tanto en los hitos prácticos como en los de investigación.

Por una parte, la existencia de un conocimiento aún fragmentario de todo el contenido de investigación impide una temporalización más detallada de sus hitos. Por otra parte las posibles dificultades técnicas pueden distorsionar seriamente la planificación de los hitos prácticos o incluso imposibilitar algunos.

En el primer caso, la estrategia es el ajuste continuado de los plazos según se avance en el trabajo de investigación y el ajuste de aquello que se incluirá finalmente en la memoria frente a lo que se descartará.

En el segundo caso, se intentarán hacer pruebas técnicas y entrenamientos casi desde el inicio, incluso aunque no se haya trabajado lo suficiente la literatura para tener el mejor conocimiento de los modelos entrenados. De esta forma se podrán subsanar problemas prácticos con suficiente margen.

El entrenamiento de alguno de los modelos puede ser algo que se demore incluso días, por lo que un conocimiento del tiempo necesario es imprescindible desde esta etapa inicial.

No se han detectado otros riesgos específicos.

5 Contenido de las PECs y de la entrega final

5.1 PEC 2

- Entrega: 20 de abril

- Contenido: seguimiento del trabajo

En esta entrega se condensará el estado del trabajo planificado hasta el momento y se ajustará la planificación si fuese necesario.

Se presentarán también los resultados de los primeros hitos logrados.

5.2 PEC 3

- Entrega: 16 de mayo
- Contenido: seguimiento del trabajo

En esta segunda entrega de seguimiento se deberán haber cubierto la mayoría de hitos del trabajo y tener bajo control las necesidades de los restantes si es que existiesen.

Se espera tener ya un borrador de redacción de algunos apartados de la memoria final.

5.3 PEC 4

- Entrega: 7 de junio
- Contenido: memoria final

Se proyecta tener un borrador final de esta memoria el día 1 y usar los días restantes para revisiones, ajustes y mejoras de su redacción.

5.4 PEC 5a

- Entrega: 12 de junio
- Contenido: presentación

6 Estructura de la memoria del TFG

En base a la plantilla proporcionada se propone la siguiente tabla de contenidos para la memoria final:

1. Resumen
2. Introducción

   (a) Contexto y justificación del Trabajo

   (b) Objetivos del Trabajo

   (c) Enfoque y método seguido

   (d) Planificación del Trabajo

   (e) Breve sumario de productos obtenidos

   (f) Breve descripción de los otros capítulos de la memoria

3. Estado del arte
4. Metodología
5. Resultados
   (a) Investigación
   (b) Producto
6. Discusión
7. Conclusiones
8. Glosario
9. Bibliografía
10. Anexos