Instituto Universitario de Investigación de Ingeniería de Aragón Universidad Zaragoza

VIVOLAB

Universidad Zaragoza 1542

XIII JORNADA DE JÓVENES INVESTIGADORES/AS DEL I3A

Uso de Sistemas *Text to Speech* (TTS) para la Síntesis de Voces Sanas y Patológicas

Santiago Rubio, Dayana Ribas, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

{s.rubio,dribas,lleida}@unizar.es

http://www.vivolab.es/

Resumen

Este trabajo investiga el comportamiento y eficacia de los **sistemas TTS** en la síntesis de **voces patológicas**. Se busca ampliar los conjuntos de datos patológicos para fortalecer la detección de tales condiciones, ofreciendo una perspectiva innovadora sobre el uso de los sistemas TTS más allá de la clonación vocal.

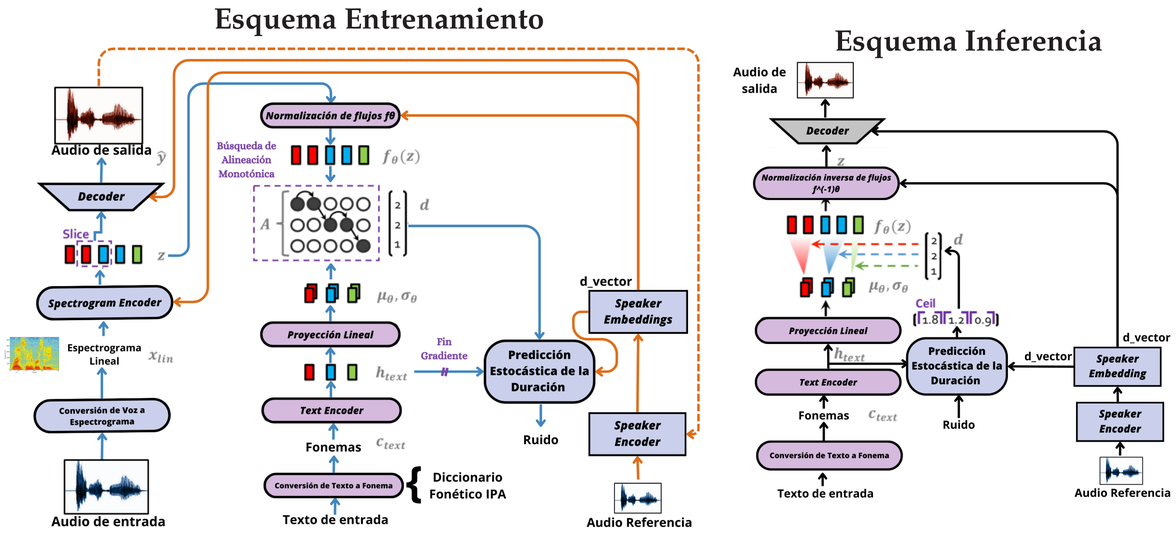
Objetivos

Este trabajo es parte del proyecto **THALENTO**, el cual desarrolla tecnologías de procesamiento del habla y lenguaje natural para estudiar trastornos de la comunicación en español. Una fase del proyecto consiste en crear una a**plicación para la detección y seguimiento automático de patologías de la voz**, destinada a asistir a los especialistas del **Hospital Clínico "Lozano Bleza"** de Zaragoza.

Sin embargo, el entrenamiento de los modelos de detección requiere una gran cantidad de audio patológico, disponiendo actualmente una cantidad limitada de este. Por lo tanto, este trabajo se enfoca en desarrollar modelos de síntesis de habla personalizados para aumentar los datos de estas voces.

Modelo: VITS

VITS se trata de un modelo *TTS multispeaker* basado en entrenamiento adversativo, inferencia variacional y normalización de flujos. El esquema del modelo se representa de la siguiente forma:



Bases de datos y Marco Experimental

Las bases de datos seleccionadas son **Librivox**, **Albayzin** y **Thalento**.

Características	Librivox	Albayzín	Thalento
Nº de oradores	1	216	12
Patología en la voz	X	X	✓
Audio transcrito	✓	✓	X
Duración audios	Entre 2 y 10 s	Entre 2 y 6 s	1 min o 3 mins
Audio/orador	53.86 h	40 seg - 12 mins	≈ 4 mins

Cada base de datos se utilizó para una etapa específica y con un objetivo particular:

1. **Entrenamiento modelo español**: Para obtener un modelo TTS en español robusto, el modelo se entrenó inicialmente con **Librivox**.
2. ***Fine-tuning* con voces sanas**: Se adapta el modelo para abarcar diversas voces sanas utilizando la base de datos **Albayzin**.
3. ***Fine-tuning* con voces patológicas**: Se ajusta el modelo para generar las 12 voces patológicas de **Thalento**.

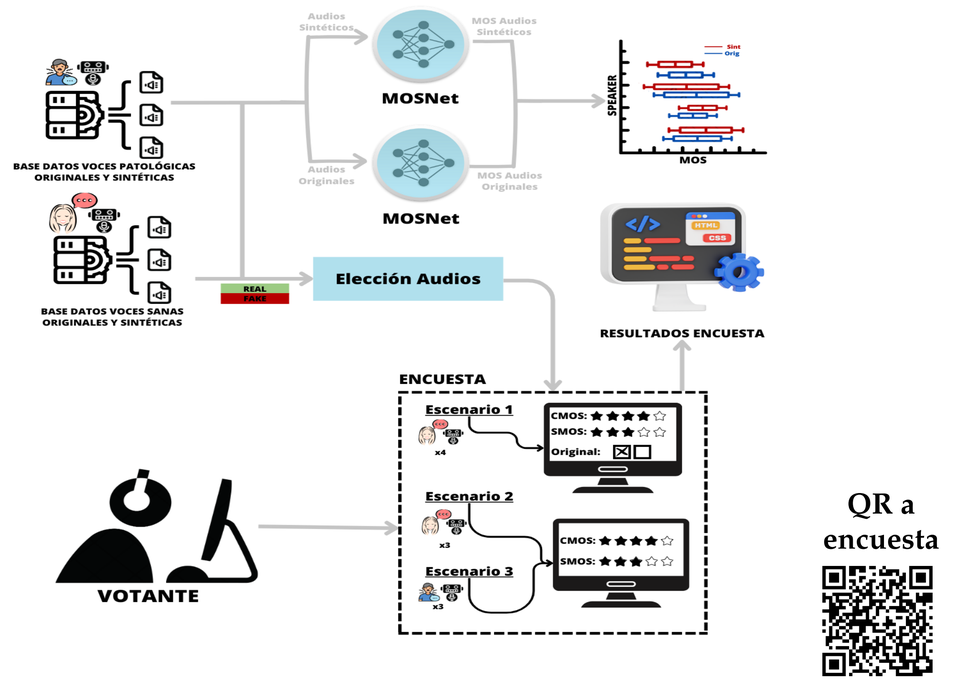
Conclusiones

- Para obtener una voz clonada de alta calidad es necesario:
  - Un **modelo inicial robusto**.
  - Diez minutos de **audios limpios, fragmentados** y con una **adecuada cobertura fonética**.
- Con una **cantidad reducida de audio**, es posible **modelar la patología de una voz**:
  - Mayor percepción de similitud en la clonación a medida que aumenta la severidad de la patología.
- Se propone un nuevo enfoque para la grabación de bases de datos patológicas:
  - Se propone **priorizar la riqueza fonética y la cantidad de audio por encima de la diversidad tonal**.

Evaluación de los resultados

Para evaluar la calidad de los audios sintetizados se utilizó la métrica **MOS** (*Mean Opinion Score*), tanto subjetivo mediante una **encuesta**, como objetivo a través del modelo ***MOSNet***. Se plantearon tres escenarios, donde se evaluó la calidad(**CMOS**) y similitud(**SMOS**) entre los audios originales y sintéticos:

1. **Evaluación de 4 voces sanas con mayor cantidad de audio**
2. **Evaluación de 216 voces sanas con diferente cantidad de audio**
3. **Evaluación de 12 voces patológicas**



QR a encuesta

Resultados

- **Primer escenario**: Valores altos de calidad y similitud con **CMOS/SMOS 3.85/5**.
- **Segundo escenario**: Calidad cerca de **3/5 sin resultados satisfactorios** en las tres voces evaluadas. La voz masculina 'ts' con 75 oraciones mostró mayor similitud, mientras que la voz femenina 'dh' fue más difícil de modelar.
- **Tercer escenario**: Pacientes con patologías severas obtuvieron mejores resultados en SMOS ($\approx 4/5$) comparados con el paciente con patología leve. Los resultados CMOS fueron inferiores debido a la baja calidad de las voces originales, pero los resultados MOS objetivo indican una **calidad similar entre audios sintéticos y originales**.

		Evaluación			Resultados	
Escenario	**Audio**	**Persona**	**Género**	**Patología**	**SMOS**	**CMOS**
1	0001	*aa*	*F*	*Sano*	**4,22**	**4,56**
	0002	*aa*	*F*	*Sano*	4, 00	4, 04
	0003	*aa*	*F*	*Sano*	4, 05	4, 04
1	0001	*mb*	*M*	*Sano*	4, 35	**4,47**
	0002	*mb*	*M*	*Sano*	**4,53**	4, 41
	0003	*mb*	*M*	*Sano*	4, 00	4, 00
2	*Mix*	*oi*	*M*	*Sano*	3, 25	3, 00
		dh	*F*	*Sano*	2, 94	3, 13
		ts	*M*	*Sano*	**3,39**	**3,87**
3	*Mix*	*TVD D* 0001	*F*	*Severa*	3, 94	**3,83**
		TVD D 0002	*M*	*Severa*	**4,00**	3, 67
		TVD D 0005	*F*	*Leve*	3, 40	3, 60

MOS objetivos de la *MOSNet*