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

Santiago Rubio Felipo, Dayana Ribas González, Eduardo Lleida Solano

Voice input Voice output Laboratory (ViVoLab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: s.rubio@unizar.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.

Introducción

Las tareas de detección de patologías en la voz siempre se han encontrado limitadas por la escasa cantidad de datos disponibles. Para intentar mejorar estos resultados otros trabajos utilizan técnicas de *data augmentation* como "*Mixup"* o "*SMOTE"* [1].

Los sistemas de síntesis de voz ("*Text to Speech*") buscan generar voces naturales e inteligibles mediante texto. Estos sistemas han estado caracterizados por la gran cantidad de audio necesitado para ser entrenados. Por contra, la cantidad de datos sobre voces patológicas en español es escasa. A esto hay que añadir que las bases de datos patológicas suelen priorizar la diversidad tonal y calidad acústica, frente a la cantidad de audio.

Afortunadamente, en los últimos años se han desarrollado modelos abiertos que pueden pre-entrenarse con bases de datos extensas y luego adaptarse ("*finetune*") para una tarea más específica con una menor cantidad de audio.

Marco Experimental

Para el desarrollo del experimento se escogió el modelo VITS [2]. VITS se trata de un modelo TTS *multispeaker* basado en entrenamiento adversativo, inferencia variacional y normalización de flujos. En el artículo *YourTTS* [3], se comprueba la alta calidad de los audios generados y la capacidad de adaptación a otras tareas con una cantidad reducida de recursos.

Respecto a las bases de datos utilizadas fueron tres: Librivox, Albayzin [4] y Thalento. Librivox, con un solo orador y numerosas horas de audio transcrito, se utilizó para entrenar un modelo en español. Albayzin incluye grabaciones de 216 personas con muchos audios cortos transcritos, sumando entre 40 segundos y 12 minutos de duración por persona. Esta base de datos fue empleada para evaluar la clonación en voces sanas. Thalento, creada por el grupo ViVoLab de la Universidad de Zaragoza, contiene grabaciones de 196 personas, divididas entre voces sanas y patológicas. Estas grabaciones fueron organizadas y guiadas por protocolos médicos de otorrinolaringólogos y logopedas del Hospital Clínico de Zaragoza. De estas, solo 29 voces disponían de suficiente audio para el *finetune* del modelo, incluyendo 12 voces patológicas con Edema de Reinke (11 mujeres y 1 hombre).

Se plantearon tres escenarios para los resultados: *fine-tuning* con las cuatro personas de Albayzin con mayor cantidad de audio, otro con las 216 personas de Albayzin, y otro con las doce voces patológicas de Thalento. El primer escenario evalúa el potencial de síntesis con una cantidad media de audio y gran riqueza fonética; el segundo, los límites del sistema para manejar múltiples personas; y el tercero, la adaptación a voces patológicas.

A la hora de evaluar métricas en sistemas de síntesis de voz (TTS) es difícil cuantificar objetivamente la calidad del audio generado. El método de evaluación más comúnmente utilizado son las *Mean Opinion Score* (MOS), donde humanos realizan un análisis subjetivo de la calidad del audio, evaluándolos en una escala de 0 a 5.

Un problema inherente al MOS es la dificultad de reunir un número suficiente de votantes asegurando que los resultados no introduzcan variabilidad. Para abordar este desafío, se utilizaron dos enfoques: una MOS subjetiva obtenida mediante una encuesta realizada a 22 votantes y una MOS objetiva usando *MOSNet* [5], modelo que estima la calidad del audio.

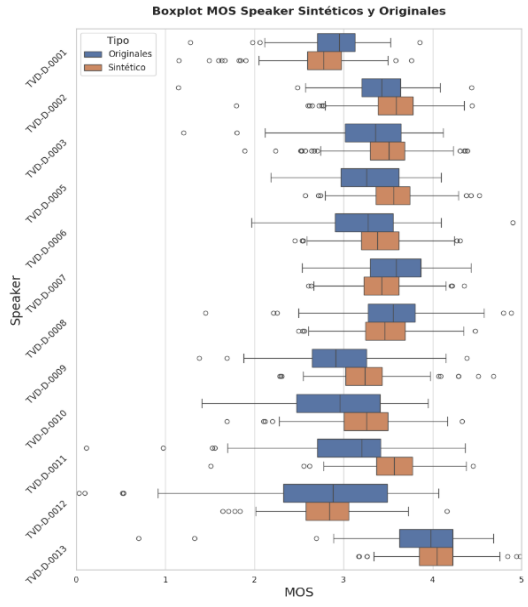
La encuesta se dividió en dos partes. En la primera, se compararon tres audios originales, excluidos del entrenamiento, con sus versiones sintéticas, evaluando la similitud (SMOS) y la calidad (CMOS). En la segunda, se usó todo el audio disponible para el entrenamiento y se solicitó calcular la métrica MOS mediante el promedio de calidad de varios audios sintéticos frente al original. Además, mediante *MOSNet* se evaluó objetivamente la calidad MOS de 250 audios sintetizados de las 12 voces patológicas y los audios originales de esas voces.

Resultados

Los resultados de la encuesta se muestran en la Tabla 1 y los resultados objetivos en la Figura 1.En el primer escenario se obtuvieron altos valores de similitud y calidad, con CMOS/SMOS ≥ 3.85/5. En el segundo escenario, los valores fueron inferiores, con resultados de calidad cercanos a 3/5 y ningún resultado satisfactorio en las tres voces evaluadas. En SMOS, la voz masculina 'ts' con 75 oraciones mostró mayor similitud, mientras que 'dh', una voz femenina, fue más difícil de modelar debido a su timbre y rango vocal.

En el tercer escenario, los pacientes 'TVD-D-0001' y 'TVD-D-0002', con patologías severas obtuvieron mejores resultados en SMOS (≈4/5) comparados con el paciente con patología leve. Sin embargo, los resultados CMOS fueron inferiores debido a la baja calidad de las voces originales. Aun así, los resultados de MOS objetivo muestran que, en los audios sintéticos y originales de cada persona, el MOS es similar. Esto indica que el audio generado puede tener una calidad similar al original.

Figura 2: Resultados *MOSNet*



Conclusiones y líneas futuras

De los dos primeros escenarios se concluye que para obtener con una voz clonada de buena calidad es necesario disponer un modelo de partida consistente y 10 minutos de audios limpios, fragmentados y con una buena cobertura fonética.

Del tercer escenario se infiere que con una cantidad baja de audio se puede llegar a modelar la patología de una voz, con una percepción de mayor similitud conforme aumenta la severidad de la patología. Esto nos permite poder generar una mayor cantidad de audio de voces patológicas para futuras tareas.

Respecto a las líneas futuras del trabajo, se pretende utilizar los resultados para la mejora del rendimiento de los detectores de patologías.

Referencias

[1]. RIBAS, Dayana, et al. On the Problem of Data Availability in Automatic Voice Disorder Detection. SCITEPRESS - Science and Technology Publications, 2023.

[2]. KIM, Jaehyeon, KONG,J.SON,J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.

[3]. CASANOVA, Edresson, et al. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone.

[4]. CASACUBERTA NOLLA, Francisco, et al. Desarrollo De Corpus Para Investigación En Tecnologías Del Habla (Albayzin). Sociedad Española para el Procesamiento del Lenguaje Natural, 1992. ISSN 1135-5948.

[5]. 1. LO, Chen-Chou, et al. MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion.

Tabla 1: Resultados de la encuesta

Escenario 1 Albayzin : SMOS				
Audio/Speak.	aa(F)	ab(F)	ma(M)	mb(M)
Audio 0001	4,22	4,35	4,07	4,35
Audio 0002	4,00	4,00	4,13	4,53
Audio 0003	4,05	4,10	3,87	4,00
Escenario 1 Albayzin : CMOS				
Audio/Speak.	aa(F)	ab(F)	ma(M)	mb(M)
Audio 0001	4,56	4,40	4,36	4,47
Audio 0002	4,04	4,15	4,07	4,41
Audio 0003	4,04	3,90	3,87	4,00
Escenario 2 Albayzin 216 Speakers : MOS				
Persona	Género	Nºorac.	CMOS	SMOS
oi	Masc.	25	3,00	3,25
dh	Fem.	50	2,94	3,13
ts	Masc.	75	3,39	3,87
Escenario 3 THALENTO : MOS				
Paciente	Género	Sever	CMOS	SMOS
TVD-D-0001	Fem.	Severa	3,83	3,94
TVD-D-0002	Masc	Severa	3,67	4,00
TVD-D-0005	Fem.	Leve	3,60	3,40