Análisis sistémico para comparar voz natural contra sintetizada

Gallegos Castillo Alexis, Marcial Margarito Sánchez Sánchez, Flavio Arturo Domínguez Pacheco
Institución: Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Zacatenco, Sección de Posgrado e Investigación
Email: cafaflip@gmail.com; mmsanchezs@ipn.mx
URL ORCID: https://orcid.org/0009-0000-0719-9065; https://orcid.org/0009-0009-0709-6651

Resumen-

En la comparación entre la voz natural y la voz sintetizada, se utilizó un análisis sistémico para identificar los aspectos en los que difieren. Al tomar como sistema los dos tipos de voz, se busca observar tanto los aspectos en los que coinciden como aquellos en los que se diferencian. Una de las principales diferencias radica en que la voz sintetizada no presenta una prosodia que suene aceptable para el oído humano, así como tampoco muestra una aleatoriedad en varias características que sí están presentes en la voz humana. Por ello, al examinar mediante el análisis sistémico, se pueden señalar las áreas de mejora y, en estas, realizar una intervención en el sistema o subsistemas. Por ejemplo, existen oportunidades de mejora al considerar la aleatoriedad presente en la voz humana, especialmente en factores como la expresividad y la transmisión de emociones.
***Palabras Clave* — voz sintetizada, voz natural, análisis sistémico, sistémica.**

Abstract-

In the comparison between natural and synthesized voices, a systemic analysis was used to identify the aspects in which they differ. By considering the two types of voice as a system, the aim is to observe both the aspects in which they coincide and those in which they differ. One of the main differences is that the synthesized voice does not present a prosody that sounds acceptable to the human ear, nor does it display randomness in several characteristics that are present in the human voice. Therefore, when examining through systemic analysis, areas for improvement can be identified and, in these, interventions can be made in the system or subsystems. For example, there are opportunities for improvement when considering the randomness present in the human voice, especially in factors such as expressiveness and the transmission of emotions.

***Keywords* — synthesized voice, natural voice, systemic analysis, systems thinking.**

I. INTRODUCCIÓN

La voz es un fenómeno fisiológico que se ha ido desarrollando respecto a las necesidades de la humanidad, donde se ha perfeccionado el uso de todos los elementos involucrados para refinar y perfeccionar la emisión de sonidos. Ayudando a hacer la comunicación más fácil y de cierto modo llevarla a otro nivel respondiendo a las demandas que requiere. Haciendo esta reflexión y tomando en cuenta que el desarrollo de la comunicación que usa el aparato fonador como principal elemento, vemos que parte desde el hombre primitivo, el cual usaba sonidos simples, como gruñidos, gritos o combinación de sonidos y gestos corporales. Como punto de partida se tiene la forma primitiva de comunicación, para lo que hoy consideramos una evolución y uso más complejo de la voz en la que se generan sonidos más complejos que ayudan a establecer una mejor comunicación, esto ayudado en parte del mayor uso de todas las partes del aparato fonador para emitir sonidos específicos [1-3].

La complejidad que llevó a tener el uso de la voz como se tiene hoy en día, comprende a partir de la simple consideración de que el aparato fonador tiene una variedad de elementos a considerar que van desde el sistema respiratorio (pulmones, bronquios y tráquea), sistema fonatorio (laringe y cuerdas vocales), sistema articular (cavidad bucal, cavidad nasal y laringe), lleva consigo una complejidad muy grande el querer imitar la voz. Esto viendo que respecto a los sistemas mencionados y sus componentes traen consigo una gran cantidad de factores que se pueden considerar como variables y estas se deben de analizar de manera aislada, para después ver cómo trabajan en la interacción de otros elementos del sistema en general que se está considerando: la voz (natural y sintetizada) [4-5].

De acuerdo con lo anterior, se puede tomar en cuenta un procedimiento que permite sintetizar la voz consiste de manera general en [6]:

1. Análisis del texto: se refiere a que el texto ingresado va a pasar a convertirse en una representación fonética.
2. Prosodia: se determina el ritmo, entonación y el énfasis

3. Síntesis de voz: Se genera una señal correspondiente a lo propuesto por los puntos anteriores
4. Post-procesamiento: se ajustan las variables para generar una señal
5. Salida: se manda la señal a algún tipo de altavoz

El proceso de sintetización de la voz es complejo, ya que mediante procedimientos matemáticos basados en diferentes algoritmos, modelos estadísticos, síntesis neuronal, entre otros. Se puede hacer la emisión de sonido que es similar a un grado considerable con la voz natural. Lo que lleva a pensar que se están contemplando y manipulando variables que implican la generación de la voz natural. Cabe mencionar que también se puede presentar un método en algunos casos que consiste en la concatenación, proceso que consiste en la pregrabación de la voz humana en fonemas, sílabas o palabras. Por lo cual, es un proceso que también se considera como sintetización, pero aquí se trata de hacer referencia al procedimiento que parte de la generación del sonido mediante los métodos mencionados. [7]

En este sentido se busca hacer uso de la sistémica para que esto ayude a identificar las áreas de mejora donde puede actuar la sintetización de la voz.

II. Metodología/Desarrollo

Contemplar el uso de una metodología como la sistémica conlleva el abordar problemas con herramientas o perspectivas como lo son el holismo, la interdependencia, la jerarquía de los elementos del sistema o la misma jerarquía de los subsistemas, la existencia de una posible retroalimentación y a partir de eso una adaptabilidad para que el sistema cambie respecto a las necesidades expuestas por este tipo de análisis [8].
Aplicando la metodología se usarían los siguientes pasos:

- Definición del problema
- Identificación de elementos del sistema
- Análisis de interrelaciones
- Modelación del sistema
- Evaluación
- Toma de decisiones

Para lo que se planea hacer este análisis y comparar los procesos y elementos de la generación de la voz natural y la voz sintetizada:

Voz natural

Definición del problema: se puede hacer una interpretación de problema como una delimitación a este. En este caso se contempla la generación de voz de forma natural y por lo cual, se deben de tomar en cuenta a los elementos y sistemas involucrados. Partiendo de que las características específicas que tienen las partes mencionadas se deben de tomar como variables dentro del sistema general que se está mencionando.

Identificación de los elementos del sistema: los principales elementos del sistema de generación de voz son los sistemas antes mencionados (respiratorio, fonatorio, articular) y algunos elementos más específicos a analizar, como lo sería el lenguaje o idioma utilizado, la persona en específico que es emisora, el mensaje que vaya a dar, la prosodia, etc. Elementos que el humano ha logrado utilizar de manera óptima para tener un sistema de comunicación auditiva.

Análisis de interrelaciones: las interrelaciones se pueden considerar principalmente entre los subsistemas de la voz antes mencionados, pero una de las primeras o más importantes interrelaciones que se deben contemplar son las que se tienen con los otros humanos y siendo más específicos se hace referencia al sistema auditivo, ya que, a partir de la percepción de este, determina si el mensaje emitido se comprende o cumple la función que esperaba el emisor al transmitir el sonido. [9]

Modelación del sistema: para poder interpretar la voz natural como sistema, se debe de interpretar el fenómeno como un sistema.

1. Elementos: principalmente el sistema respiratorio, fonatorio y articular. Algunos otros elementos se mencionaron anteriormente
2. Interrelaciones: se muestra al ver la dependencia entre los subsistemas involucrados para la generación de voz y como la variación en estos, afecta directamente al resultado emitido.
3. Límites: se plantea solamente observar la generación de voz, sin abarcar directamente aspectos como ver el mensaje, el idioma, el sistema auditivo y el entorno físico o social que involucra el proceso de la emisión de la voz.
4. Ambiente: la generación de la voz depende totalmente del entorno del emisor y de las características propias que posee la persona.
5. Entradas: para interpretarlo de manera simple, la entras que se consideran son las ideas que se buscan expresar mediante la voz
6. Procesos: al pasar por cada subsistema, se le da una caracterización específica a la voz
7. Salidas: se considera al sonido de la voz como salida
8. Retroalimentación: el cambio de emisión de ideas que se quieran transmitir dado por el emisor.
9. Objetivo: determinar la similitud de procesos de la voz natural con la voz sintetizada
10. Jerarquía: se considera como el sistema principal al sistema emisor de la voz y como subsistemas el fonador, el respiratorio y el articular.

.

Evaluación: En este sistema se propone que mediante un proceso llamado inteligibilidad del habla se evalúe la calidad del sonido emitido respecto a la emisión de voz.

Este tipo de evaluación contempla varios factores como lo son: el hablante, el oyente y el entorno. Para este tipo de análisis solamente se contemplan factores del hablante que incluyen a la articulación, la velocidad de emisión, volumen, acento o dialecto y fluidez. Estas características se consideran importantes al momento en el que un hablante emite voz para que su mensaje sea bien recibido y manejar estas variables en porcentajes elevados, permite que exista una naturalidad entre hablantes.

Toma de decisiones: El ver la variación de las diferentes características del hablante permite tener diferentes consideraciones respecto al análisis de voz, al tener un análisis mejor de los sistemas y factores que están involucrados en la emisión de voz natural, ayuda a que estas mismas características sean consideradas al momento de sintetizar la voz.

Voz sintetizada

Definición de problema: dentro del análisis realizado, al comparar la voz sintetizada con la natural, se presentan diferentes problemas como que la sintetización genera una carencia de sutiles variaciones en el tono, ritmo y volumen, no presenta una adaptabilidad, algunas que cumplen con esos aspectos dependen de grabaciones de voz humana, pero el principal problema es que para el escucha puede parecer carente de emociones.
Elementos del sistema: algunos de sus elementos principales parten de la idea que se busca expresar mediante esta herramienta, en este caso es una entrada de texto. Lo siguiente es el procesamiento de texto seguido por un modelamiento lingüístico, una generación de prosodia y se genera una síntesis de sonido, que en ciertos casos se puede tener un post procesamiento que pasa a ser amplificado por algún dispositivo. En la actualidad este proceso es apoyado por inteligencia artificial que va aprendiendo y se retroalimenta

Análisis de interrelaciones: los elementos deben de tener una interrelación para que se lleve a cabo la síntesis.

Modelación de sistema: se considera a grandes rasgos y sin ser específicos, que para que se pueda interpretar como sistema se requiere una entrada que este caso es la entrada de texto, seguida por los procesos para generación y manipulación del audio de la voz que tienen por salida una señal de audio que puede ser amplificada y como posible retroalimentación se puede considerar a la inteligencia artificial que determina que con la información que se le da de retroalimentación puede cambiar o manipular las señales de voz generadas.

Evaluación: Se puede determinar que simplemente mientras el escucha no sienta que de un modo u otro no considera que la voz generada artificialmente carece de emociones o no tiene la calidad que le agrada, no va a ser completamente óptima para todas las situaciones.

Toma de decisiones: el conocer las variables que involucran la generación de la voz natural permite en cierta manera hacer una comparación directa sobre los factores donde falla la síntesis de la voz. Por lo cual, se debe de tener consideración de que, mediante una comparación de sistemas, se pueden llegar a considerar aplicar o mejorar algoritmos de la síntesis que permitan manipular las variables para que la voz sintetizada se parezca cada vez más a la humana, en los casos donde no parte de una grabación previa.
.

III. Resultados y Discusión

La comparación directa recae en que de cierto modo a estas alturas hay elementos que no pueden ser imitables y que presentan un sinnúmero de nuevas variables como lo puede ser la prosodia, la cual es referente a los matices de las emociones humanas y esto da elementos a considerar como la melodía, la cadencia, las pausas, entre otros. Características que, al considerarse aleatorias, tienen un grado de dificultad mayor para su imitación [10]. Al tener en cuenta la aleatoriedad y la no linealidad de la comunicación humana se considera como tal que hay muchas características que permiten que sea considerado como ambiguo en el aspecto de que un mismo mensaje puede ser emitido de distintos modos y este mismo puede tener distintas interpretaciones, donde se encuentran situaciones las cuales necesitan una mejor comprensión de todas las variables presentes como por ejemplo en el caso de la ironía. [11]
Al mencionar lo anterior, se vuelve al punto de comparación de la voz natural y la sintetizada, donde podemos observar que hay grandes avances respecto a la imitación y es de gran uso popular en estos días en casos como las inteligencias artificiales o los asistentes virtuales, donde los consumidores buscan más que estos suenen lo más parecido a la voz humana y que tratan de tener un equilibrio entre los sonidos pregrabados y los sintetizados. Por lo que, se plantea que un análisis sistémico sirva como herramienta que ayude a que se cumpla este cometido, o quizá sea un punto de partida para análisis futuros en el área.[12]
Comparando cada parte de los sistemas, tenemos que:

Definición de problema: la voz humana es aleatoria en muchas características y la voz sintetizada carece de eso en muchos aspectos. La prosodia aún no se logra imitar correctamente y puede ser una de las áreas de mejora.
Elementos del sistema: aquí es donde se complican las comparaciones, ya que son procesos que buscan un mismo fin, pero uno principalmente se basa en procesos matemáticos y el otro es un proceso desarrollado por la evolución humana. Pero se pueden considerar elementos y aspectos a ser imitados como el sistema respiratorio, el cual se encarga principalmente de generar un flujo de aire que hace vibrar las cuerdas vocales y ayuda a brindar una modulación del habla de manera general.

Por otra parte, el sistema fonatorio brinda variables importantes en el habla como, por ejemplo, la producción de sonido a partir de las cuerdas vocales, control tonal, presenta la coordinación entre el sistema respiratorio y el articular, y permite que trabajen en conjunto. Otro sistema o subsistema considerado es el articular, el cual está definido por las características físicas del hablante, y este es el que presenta una gran variedad de elementos a considerar al momento de generar sonidos y emitir la voz. Ya que este al tener elementos como los puntos de articulación (labiales, dentales, etc.) y los modos de articulación (oclusivas, fricativas, etc.), cada persona tiene su propia huella sonora y al querer imitar en la síntesis de voz, se debe de tener una idea generalizada de que tipo de voz es la que se va a tratar de imitar.

Teniendo en cuenta esto, se puede proponer que los algoritmos que contemplen las características brindadas por estos sistemas o algunas otras para que la voz sintetizada se parezca cada vez más a la natural

Análisis de interrelaciones: los sistemas interactúan de maneras "similares" en el aspecto de que los dos sistemas tanto el de la voz natural y el de la sintetización parten de querer transmitir un sonido, pasan por un proceso para generarlo y terminan con un sonido. La diferencia principal se encuentra entre las interrelaciones y los procesos para generar el sonido, que son muy diferentes, permitiendo que en lo general se pueda comparar ya que sus salidas (emisión de sonido: voz), se consideran similares, pero en lo particular, los procesos para la generación no parecen estar relacionados. Uno es un proceso biológico y el otro es un proceso artificial.

Modelación de sistema: parte de la representación de los sistemas. En este caso, se busca que se pueda ver una similitud entre los sistemas.

Evaluación: se menciona de nuevo que la principal diferencia y área de mejora de la voz sintetizada parte en que, si no se trata de un sonido pregrabado, se considera que topa con que presenta carencias respecto a la voz humana.

Toma de decisiones: considerar las características de la voz humana de manera sistémica, permite que se relacione este proceso con la generación de voz sintética. Y a partir de eso, se pueden plantear áreas de mejora respecto a las carencias que se identifiquen en la síntesis de voz.

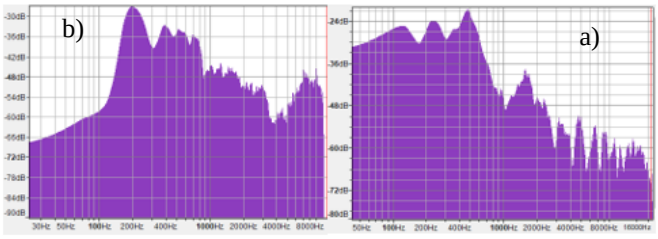

Figura 1.- Espectro de frecuencias. a) voz sintetizada. b) voz grabada (Elaboración propia)

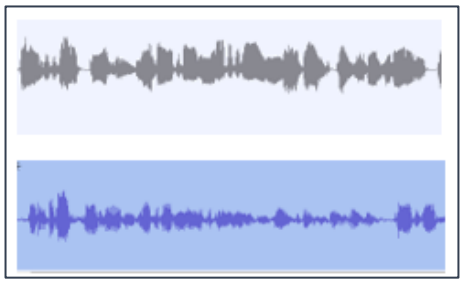

Figura 2.- Formas de onda. a) voz sintetizada. b) voz grabada (Elaboración propia)

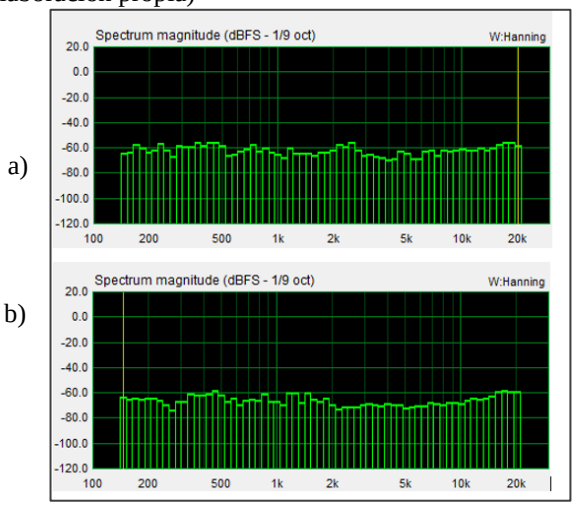

Figura 3.- Espectro de frecuencias por bandas de octava, a) voz sintetizada. b) voz grabada (Elaboración propia)

En las figuras 1-3 se puede apreciar cómo se comportan de maneras similares, pero mostrando en ciertos aspectos las voces tanto sintetizada como natural. La comparación se realizó con una voz de un sintetizador de voz (ttsmp3) y la grabación de una voz natural haciendo que digan un mensaje que tiene variedad de sonidos. El mensaje es "El rápido zorro marrón salta sobre el perro perezoso mientras las aves cantan suaves melodías al amanecer".

Lo más notable a mencionar es que la voz natural en la Figura 1 permite ver cómo tiene más presencia y variación de frecuencias en comparación con la sintetizada. En la Figura 2, se puede ver como la forma de onda de la voz natural muestra como hay más variedad en la intensidad de la voz. En la Figura 3 se ve como hay comportamiento similar y que en ciertos rangos de frecuencias se puede ver que hay variaciones relevantes, pero mediante el espectro de frecuencias representado por bandas de octava, no permite ver de manera tan explícita como las otras representaciones, aunque esta representación da preámbulo a ver las áreas donde se presentan más diferencias.

IV. Conclusiones

Dentro del proceso de análisis sistémico, se considera una visión holística del problema de la generación de voz considerando a los tipos de generación como sistemas, donde coinciden principalmente en que sus salidas son consideradas como sonido o voz. Y al tener en cuenta esto, se trata de ver como dos sistemas diferentes que carecen de similitudes, se puede tratar de hacer una semejanza de procesos, para que se puedan encontrar áreas de mejora y que estas se puedan trabajar para que los procesos, que, aunque no sean similares, den como resultado una salida de sonido similar, o por lo menos que se pueda interpretar cada vez de manera más humana o que no choque directamente con la percepción del escucha.

AGRADECIMIENTOS

Los autores agradecen a las personas involucradas por el apoyo brindado.

DECLARACIÓN ÉTICA

Los autores garantizan que este trabajo cumple con los más altos estándares éticos y de integridad científica. Se asegura la veracidad de los datos, el uso de metodologías válidas y la ausencia de prácticas como plagio o falsificación. En el marco de esta investigación, que realiza una comparación entre la voz natural y la voz sintetizada, se aseguró que el análisis y la recopilación de datos se llevaron a cabo de manera transparente y rigurosa, respetando los principios éticos en el uso de tecnologías de síntesis de voz. Los autores se comprometen a cumplir las normas éticas del congreso y a garantizar que los resultados presentados contribuyan de manera responsable al avance del conocimiento en este campo.

REFERENCIAS

[1]. Lieberman, P. (2006) Toward an Evolutionary Biology of Language. Cambridge, MA: Harvard University Press. doi:10.2307/j.ctv22jnsvv.

[2]. Johansson, S. (2005) Origins of Language: Constraints on Hypotheses. Amsterdam: John Benjamins Publishing Company. doi:10.1075/celcr.5.

[3]. Titze, I. R. (2008) 'The human instrument', Scientific American, 298(1), pp. 94-101. doi:10.1038/scientificamerican0108-94.

[4]. Rabiner, L. y Schafer, R. (2011) Theory and Applications of Digital Speech Processing. 1st edn. Upper Saddle River, NJ: Pearson.

[5]. Kent, R. D. y Read, C. (2002) The Acoustic Analysis of Speech. 2nd edn. Albany, NY: Thomson Learning.

[6]. Williams, C. E. y Stevens, K. N. (1972) 'Emotions and speech: some acoustical correlates', The Journal of the Acoustical Society of America, 52(4), pp. 1238-1250. doi:10.1121/1.1913238.

[7]. Marulanda Torres, L. (2020) Protocolos y herramientas para la construcción de una síntesis de voz. Pereira: Universidad Tecnológica de Pereira. Disponible en: https://hdl.handle.net/11059/12201.

[8]. Von Bertalanffy, L. et al. (1996) Teoría general de los sistemas [en línea]. Disponible en: https://www.academia.edu/download/59471390/TGS_Bertalanffy20190531-130081-rt2nka.pdf [Consultado: 4 de octubre de 2023].

[9]. Bubnova, T. (2006) 'Voz, sentido y diálogo en Bajtín', Acta poética, 27(1), pp. 97-114.

[10]. Ruiz, S., Miranda, E., Herlein, M., Etchart, G. y Alvez, C. E. (2017) 'Análisis comparativo de distintas toolkits para el reconocimiento biométrico de personas mediante voz'. En: XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires).

[11]. Balordi, A. E. (1997) 'El concepto de ironía: de tropo a ambigüedad argumentativa'. En: Homenaje al profesor J. Cantera. Madrid: Servicio de Publicaciones Complutenses, pp. 451-461.

[12]. Franganillo, J. (2023) 'La inteligencia artificial generativa y su impacto en la creación de contenidos mediáticos', methaodos. Revista de ciencias sociales, 11(2), p. 10.

DECLARACIÓN DE NO CONFLICTO DE INTERÉS

Los autores declaran que no tienen ningún conflicto de intereses que pueda influir en la objetividad, integridad o interpretación de los resultados presentados en este trabajo. No existen relaciones financieras, personales o profesionales con terceros que puedan ser percibidas como un conflicto de intereses en relación con esta investigación. Este trabajo se ha realizado de manera independiente y sin influencias externas que pudieran comprometer su contenido.