Comparación de Calidad de Recuperación de Señales de Voz Sintética y Voces Naturales en Idioma Español Mediante Muestreo Compresivo

Mauro Leonel Furer, Ricardo Tomás Ferreyra

Universidad Nacional de Córdoba - Facultad de Ciencias Exáctas Físicas y Naturales, Av. Vélez Sarsfield 299, Córdoba - Argentina
mauro.furer@mi.unc.edu.ar,ricardo.tomas.ferreyra@unc.edu.ar

Abstract. En este trabajo se analiza la calidad de recuperación de señales de voces naturales y sintética en idioma Español utilizando técnicas de muestreo compresivo. Se tomaron muestras de voces masculinas de un adulto mayor, un adulto, un niño y una voz sintética. Se utilizaron distintas técnicas de representación dispersa y se evaluaron métricas de calidad, como la relación señal-ruido, el error cuadrático medio y el coeficiente de correlación, para comparar las señales originales con las recuperadas para distintos factores de compresión. Naturalmente, los resultados muestran que la calidad de recuperación de las señales de las voces naturales es superior a la de la sintética, como así también el número de muestras tienen un impacto significativo en ese aspecto. Sin embargo, la herramienta desarrollada muestra claramente las propiedades y limitaciones de la voz sintética recuperada.

Keywords: Muestreo compresivo · Voz Sintética.

1 Introducción

El paradigma del muestreo compresivo (CS, compressed sensing) establece que es posible recuperar ciertas señales e imágenes partiendo de una cantidad menor de muestras que las convencionales [1] establecidas por el teorema del muestreo de Nyquist.

El aporte de este trabajo es utilizando un conjunto caracterizado por frases pronunciadas en idioma Español, por voces masculinas provenientes de: un niño, un adulto, un adulto mayor y una voz sintetizada por el servicio de Microsoft speech, crear una herramienta para recuperar las señales utilizando CS y evaluar el desempeño de la recuperación.

2 Materiales y Métodos

Cada participante pronunció (Frase A): “Hace doce grados con sol”. (Frase B): “Son las veintrés horas cincuenta y siete minutos”.

El problema de optimización convexa se expresó como en [3]:

$$\hat{s} = \min \|s\|_{L1} \quad s.t \quad y = \Theta s = \Phi \Psi^{-1} s \quad (1)$$

Donde $\Phi \in C^{m \times n}$ con $m << n$ es la matriz de muestreo y $\Psi^{-1} \in C^{n \times n}$ representa la transformada inversa de Fourier discreta (IDFT, inverse DFT) o también a la transformada inversa del coseno (IDCT, inverse DCT).

3 Experimentos y Resultados

Las señales fueron segmentadas en ventanas de 256 muestras usando ventana Hann [2], de 32 [ms] de duración y solapamiento de 50 %. La recuperación se realizó resolviendo (1) utilizando Basis Pursuit (BP)[4] y solver [5], obteniendose $\hat{x} = \Psi^{-1}\hat{s}$. Se evaluó la similitud en cada caso entre la señal original y la recuperada.

3.1 Recuperación usando IDCT

En este caso se utilizó la matriz de la IDCT en el problema (1), obteniéndose:

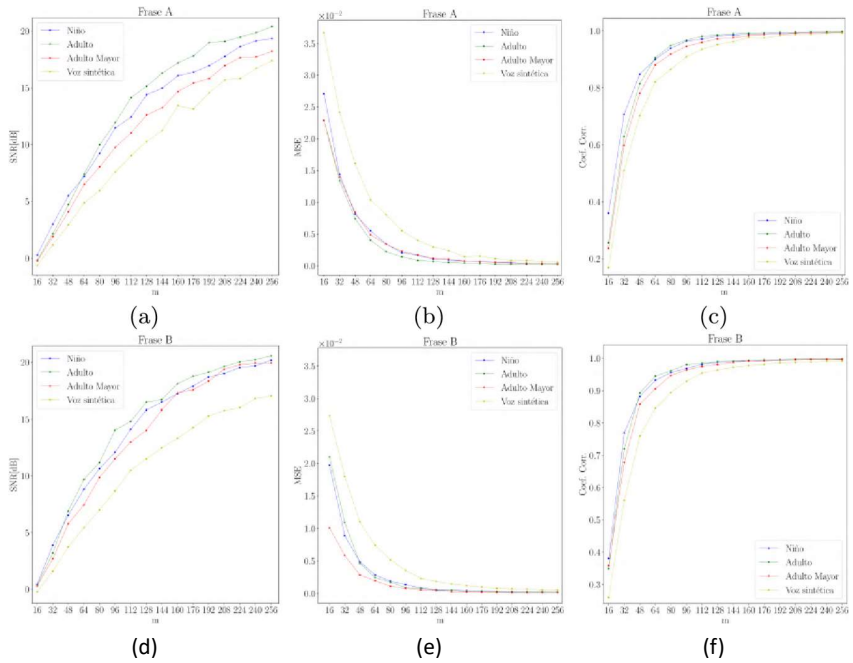

Fig.1. Recuperación mediante IDCT: (a) SNR Frase A. (b) MSE Frase A. (c) CC Frase A. (d) SNR Frase B. (e) MSE Frase B. (f) CC Frase A.

Tomando un valor "m" de 128 muestras que representa una compresión del 50 % y observando el valor de SNR se tiene de manera ordenada de mayor a menor.

Frase A: Adulto 15.14 [dB], Niño 14.39 [dB], Adulto mayor 12.60 [dB] y Voz sintética 10.26 [dB]. Frase B: Adulto 16.47 [dB], Niño 15.80 [dB], Adulto mayor 14.02

[dB] y Voz sintética 11.49 [dB]. El mejor desempeño con valores de Relación señal-ruido (SNR, signal to noise ratio), Error cuadrático medio (MSE, mean squared error) y Coeficiente de correlación (CC). lo tiene la recuperación de la voz del Adulto. La Voz sintética tiene los peores valoresde SNR, MSE y CC. Este mismo comportamiento puede observarse para compresiones mayores del 37.5 % que corresponden a 96 muestras o más.

3.2 Recuperacion usando IDFT

En este caso se utilizó particularmente la implementación de la transformada rápida de Fourier inversa (IFFT, inverse fast Fourier transform) (1), obteniéndose:

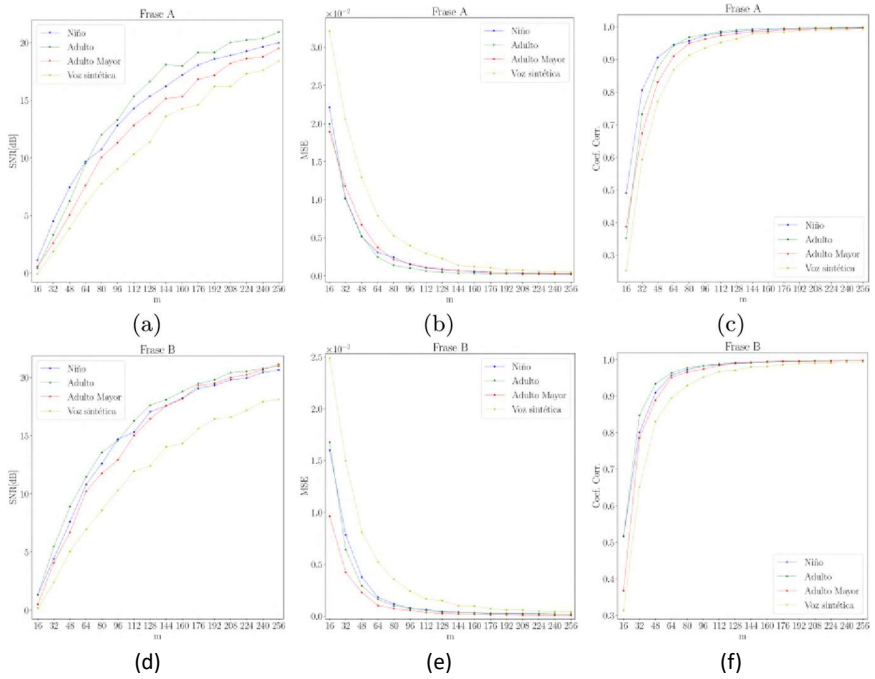

Fig.2. Recuperación mediante IFFT: (a) SNR Frase A. (b) MSE Frase A. (c) CC Frase A. (d) SNR Frase B. (e) MSE Frase B. (f) CC Frase A.

Tomando un valor “m” de 128 muestras que representa una compresión del 50 % y observando el valor de SNR se tiene de manera ordenada de mayor a menor. Frase A: Adulto 16.62 [dB], Niño 15.35 [dB], Adulto mayor 13.87 [dB] y Voz sintética 11.38 [dB]. Frase B: Adulto 17.63 [dB], Niño 17.06 [dB], Adulto mayor 16.46 [dB] y Voz sintética 12.40 [dB]. Por lo que el mejor desempeño con valores SNR, MSE y CC lo tiene la recuperación de la voz del Adulto. La Voz sintética tiene los peores valores de

SNR, MSE y CC. Este comportamiento puede observarse prácticamente para compresiones mayores del 25 % (64 muestras o más).

4 Conclusiones

En este trabajo se aplicó la herramienta desarrollada a las señales de voz en idioma español caracterizadas por voces masculinas de distintos rangos etarios y también por una voz masculina sintética del servicio Microsoft Speech. Se aplicó el algoritmo implementado a las señales mencionadas y la herramienta logró recuperar las señales de manera correcta.

Se obtuvieron mejores resultados con la transformada IDFT que con la IDCT según los indicadores utilizados. Las voces de origen humano obtuvieron mejores valores de similitud que la voz sintética, la cual tuvo los peores resultados.

References

1. Candes, E. J., Wakin M. B.: An Introduction To Compressive Sampling.In: IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 21-30. (2008). https://doi.org/10.1109/MSP.2007.914731
2. Rabiner, L. R., Schafer, R. W.: Theory and Applications of Digital Speech Processing. 1st edn. Pearson, Upper Saddle River (2011)
3. Candes, E. J., Tao, T.: Decoding by linear programming. In IEEE Transactions on Information Theory, vol. 51, no. 12, pp. 4203-4215. (2005). https://doi.org/10.1109/TIT.2005.858979
4. Chen, S. S., Donoho D. L., Saunders, M. A.: Atomic Decomposition by BasisPursuit. SIAM Journal on Scientific Computing, vol. 20, no. 1, pp. 33–61 (1998), https://doi.org/10.1137/S1064827596304010
5. Domahidi, A., Chu E. K.-W., Boyd, S. P.: ECOS: An SOCP solver for embeddedsystems. 2013 European Control Conference (ECC), pp. 3071–3076. (2013)