Conversión de Voz aplicada a personas con vocalización disminuida

Alejandro J. Uriz, Pablo D. Agüero, Juan C. Tulli, Jorge Castiñeira y Esteban L. González
Facultad de Ingeniería, Universidad Nacional de Mar del Plata, ajuriz@conicet.gov.ar, {pdaguero,jctulli}@fi.mdp.edu.ar, 0223-481-6600, Juan B. Justo 4302 (7600) Mar del Plata

***Resumen*— Los sistemas de conversión de voz (VC) cumplen la función de transformar la voz de un locutor en la de otro locutor específico. Estos sistemas constan de dos etapas: una de entrenamiento, donde el sistema adquiere información acerca de los locutores fuente y destino, y una segunda etapa que realiza la conversión de voz. La conversión de voz tiene muchas aplicaciones, desde el entretenimiento hasta la medicina.**

Estos sistemas, por ejemplo, permiten que una persona que ha perdido su voz a causa de alguna enfermedad en su sistema fonatorio, (por ejemplo un cáncer tiroideo), pueda reemplazarla a través de un dispositivo de asistencia. En este marco se desarrolla este trabajo, que propone el desarrollo de un sistema de asistencia que permita a una persona que ha perdido el habla, emularla con la misma voz que tenía antes de sufrir la enfermedad. Si bien existen sistemas de síntesis de Texto a Voz (TTS) que permiten comunicarse a una persona que ha perdido su capacidad de hablar, estos tienen el problema que la identidad de la voz sintetizada no es la original de la persona, sino una voz que depende del programa de síntesis utilizado, la cual además suele ser percibida como artificial. El trabajo aquí propuesto aborda este problema sintetizando una versión muy aproximada a la voz original del paciente, tratando de reducir de esta forma el trauma ocasionado por la pérdida del habla. Esto se logra en base a novedosas técnicas de VC, las cuales permiten superar la relación calidad-identidad de otros sistemas existentes en la literatura. Por último, se desarrolla un sistema de texto a voz que implementa este modelo.

***Palabras clave*— Conversión de Voz, Síntesis texto a voz, discapacidades de habla.**

I. Introducción

La Conversión de Voz (VC) es una técnica de procesamiento de señales que consiste en tomar la señal de voz de un locutor llamado locutor origen o *source* y por medio de un conjunto de modificaciones transformarla en la voz de un segundo locutor específico conocido como locutor destino o *target*. La conversión de voz fue presentada por Abe [1], y se ha desarrollado en los últimos años gracias al surgimiento de computadoras mas potentes. Existen diversas técnicas de VC, pero se las puede agrupar en tres grandes grupos: el primero de ellos está compuesto por sistemas que generan la voz del locutor destino en base a selección de unidades almacenadas en una base de datos previamente grabada. Estos sistemas se caracterizan por sintetizar voces con una elevada identidad, ya que la misma es generada partiendo de datos reales del locutor destino, pero la desventaja de estos sistemas es que presentan problemas acústicos debidos al volumen de datos que se utiliza para sintetizar las voces, son ejemplos de este tipo de sistemas [1] y [2]. Los sistemas del segundo grupo están basados en algoritmos que particionan los espacios acústicos de cada uno de los locutores para luego realizar la transformación utilizando funciones estadísticas, sintetizando así una aproximación de la voz del locutor destino. En este tipo de sistemas, los resultados obtenidos no poseen una identidad tan elevada como en los métodos de selección de unidades, pero no presentan las discontinuidades acústicas de los primeros métodos. Son ejemplos de este tipo de sistemas los basados en GMM (*Gaussian Mixture Model*) [3],[4] y en KH (*K-Histogramas*) [5]. Existe un grupo de sistemas de conversión de voz que combinan las ventajas de los dos anteriores para obtener una relación identidad-calidad acústica lo mas elevada posible. Los sistemas dentro de este grupo se caracterizan por estar compuestos por dos etapas de conversión. La primera realiza una conversión por partición del espacio (utilizando un algoritmo como el GMM por ejemplo), mientras que la segunda etapa realiza la conversión mediante un sistema de selección de unidades. Este tipo de sistemas ha tenido un auge en los últimos tiempos debido a los resultados que permiten obtener. Son ejemplos de ellos [6], [7] y [8]. La conversión de voz tiene diversas aplicaciones que van desde los videojuegos hasta la medicina, siendo este último el enfoque que se desarrolla en el presente trabajo.

Existen en la Sociedad personas cuya capacidad de oralización se ve disminuida por diversos tipos de enfermedades o discapacidades y es por ello que deben recurrir a dispositivos de asistencia. Generalmente estos sistemas están compuestos por un sintetizador de voz que transforma texto en voz (TTS) [9],[10] y [11]. Sin embargo, este tipo de sistemas sólo permite que la persona se exprese mediante las voces que tiene por defecto el sistema de síntesis. Esto hace que la persona pierda su voz original, lo que puede ocasionar una serie de traumas en el locutor. Es por ello que en este trabajo se pretende desarrollar un sistema de conversión de voz que permita convertir la voz que genera el sistema TTS en una aproximación a la voz del paciente antes de sufrir el trauma.

El Trabajo se presenta con la siguiente estructura: El la Sección 2 se presenta el problema a desarrollar y se presenta el sistema sintetizador de Texto a Voz. A lo largo de la Sección 3 se estudia el sistema de conversión de voz a implementar. La Sección 4 desarrolla el modelo propuesto. Por último, en la Sección 5 se presentan las conclusiones del trabajo y los lineamientos a seguir en el futuro.

II. DESARROLLO DEL PROBLEMA

Este Trabajo tiene como objetivo desarrollar un sistema un sistema que permita convertir la voz que genera un sintetizador de Texto a Voz (TTS) en la de un locutor específico. Una representación del sistema propuesto se presenta en la Figura 1.

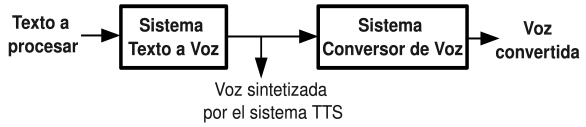

Fig. 1. Esquema del Sistema Propuesto

A. Interfaz Texto a Voz

El software de conversión texto a voz usado para la implementación del sistema descripto en la Figura 1 es el Festival [9]. Este TTS desarrollado por el Centro de Investigación de Tecnología del Habla de la Universidad de Edinburgo puede ser utilizado para la generación de nuevas voces para múltiples aplicaciones, tanto para uso comercial y no comercial en forma gratuita.

La voz generada para la conversión texto a voz en el dialecto argentino fue preparada utilizando el software disponible en el paquete FESTVOX, SPHINX y SPHINXTRAIN, tal como se describe en el capítulo 14 del documento "Building Synthetic Voices" de Alan Black y Kevin A. Lenzo.

La voz consta de un corpus de oraciones seleccionadas para lograr una gran cobertura fonética en diferentes contextos. La transcripción fonética es automática utilizando etiquetas SAMPA, y respetando las reglas correspondientes al dialecto argentino.

La calidad e inteligibilidad iniciales de la voz son adecuados, y se prevee su mejora continua a través de diferentes refinamientos en las etapas de generación de la voz. Además, se grabarán nuevas voces con una mayor cobertura tanto fonética como expresiva para mejorar la naturalidad de la voz final.

III. SISTEMA DE CONVERSIÓN DE VOZ

Dado que el sistema debe permitir al usuario comunicarse con naturalidad, se debe apuntar a desarrollar un sistema que por un lado presente una relación identidad-calidad elevada, pero que además realice el procesamiento en el menor tiempo posible. Por este motivo a lo largo de las siguientes secciones se presentan una serie de métodos de conversión de voz con el fin de seleccionar el más apropiado para la tarea a realizar.

A. Conversión de Voz Utilizando Selección de Segmentos

Debido a que el sistema a implementar requiere una voz sintetizada con una calidad acústica elevada, los métodos de selección de unidades puros (FS) [2] no son constituyen la mejor opción. Esto se debe a que este tipo de sistemas funciona seleccionando unidades almacenadas en una base de datos pregrabada. Para realizar este proceso se utiliza una minimización de la siguiente función distancia:

$$\min_{\hat{y}} \left[\sum_i d\left(x_i, x_j^{train}\right) + d\left(\hat{y}_{i-1}, \hat{y}_i\right) \right] \qquad (1)$$

Donde:

$d\left(x_i, x_j^{train}\right)$ = Distancia euclídea entre el segmento i-ésimo a convertir y el segmento j-ésimo de la base de datos de entrenamiento.

$d\left(\hat{y}_{i-1}, \hat{y}_i\right)$ = Costo de concatenación que resulta de seleccionar el segmento $\hat{y}_i$ habiendo seleccionado previamente el $\hat{y}_{i-1}$.

El conjunto de segmentos resultante de aplicar sucesivamente la Ec. 1 es el que se utiliza para sintetizar la voz convertida.

El rendimiento de este tipo de sistemas esta relacionada directamente con las características de la base de datos utilizada. Esto es debido a que si la base de datos tiene una cantidad de información reducida, presentan problemas originados por discontinuidades acústicas debidas al tamaño de la base de datos. En contraparte, si se amplía el tamaño de la base, aunque se mejora la calidad del audio resultante, el tiempo de procesamiento aumenta notablemente, haciendo que esta opción no sea viable. Un esquema de este tipo de sistemas se presenta en la Figura 3.

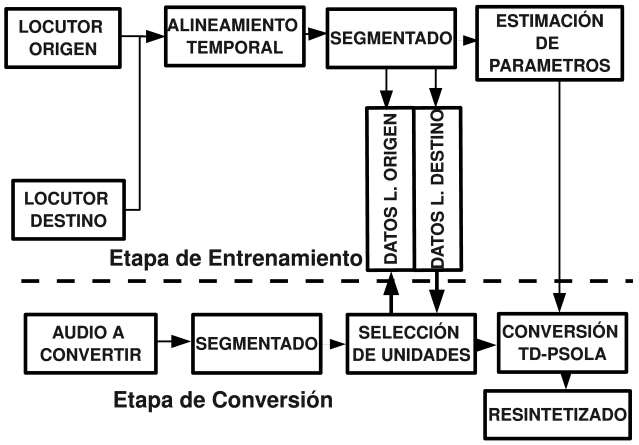

Fig. 3. Sistema de Conversión de Voz Basado en Selección de Unidades.

B. Conversión de Voz en base a partición del Espacio

Este tipo de técnicas de VC trabaja mediante técnicas de particionado del espacio o *clustering* permiten obtener voces sintetizas con una elevada calidad pero con una identidad intermedia entre los locutores fuente y destino. Una característica relevante de este tipo de sistemas es que dado que los segmentos de la señal de voz a convertir sólo son afectados por una normalización utilizando parámetros obtenidos en la etapa de entrenamiento, su implementación puede ser realizada en tiempo real.

Una técnica ampliamente usada dentro de este grupo de sistemas, es la conversión basada en Mezclas de Gausianas

o *Gaussian Mixture Models* (**GMM**) [3],[4]. Los cuales se caracterizan por agrupar los espacios acústicos de ambos locutores mediante modelos basados en N distribuciones gausianas. Una vez realizado esto, se obtiene para cada grupo una media y una varianza que permitirán transformar cada segmento del locutor origen en una estimación del correspondiente al locutor destino. La función transformación propuesta es la siguiente:

$$F(x) = \sum_{i=1}^{m} p_i \left| \mu_i^y + \sum_i^{yx} \sum_i^{xx\,-1} (x - \mu_i^x) \right| \quad (2)$$

Donde, en la Ec. 2 *F(x)* representa el segmento estimado del locutor destino que se utiliza para la conversión, *m* es la cantidad de particiones realizadas al espacio acústico de cada locutor. Por otro lado:

x = Segmento a convertir.

p_i = Probabilidad de que el segmento a convertir x, pertenezca a la clase fonética i-ésima.

μ_i^x = Vector de valores medios de los parámetros de la clase fonética i-ésima, en el espacio acústico del locutor origen.

μ_i^y = Vector de valores medios de los parámetros de la clase fonética i-ésima, en el espacio acústico del locutor destino.

$\sum_i^{xx\,-1}$ = Inversa de la matriz de covarianzas de la clase fonética i-ésima, para el espacio acústico del locutor origen.

$\sum_i^{yx}$ = Matriz de covarianzas de las clases fonéticas i-ésima, para los espacios acústicos de los locutores origen y destino.

El sistema de VC por GMM se representa en la Figura 4.

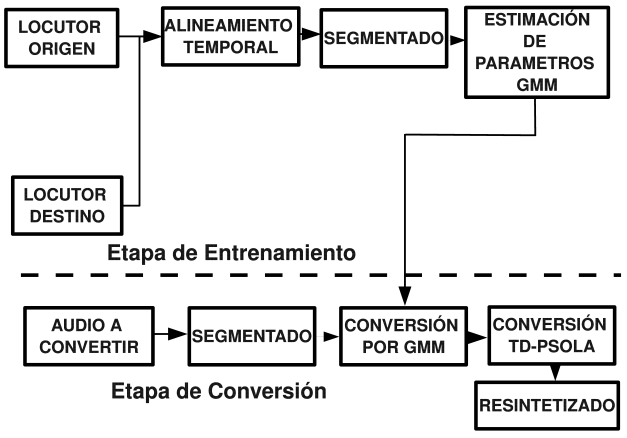

Fig. 4. Sistema de Conversión de Voz Basado en Partición del Espacio.

Una técnica de VC que supera los resultados obtenidos mediante GMM, es la conversión utilizando K-Histogramas (**KH**). Estos sistemas de conversión de voz son desarrollados en [5] y [8], y si bien son algoritmos que operan utilizando partición del espacio, tal como lo hace GMM, su filosofía de trabajo es completamente distinta, ya que en vez de considerar la información a procesar como datos numéricos, los interpreta como datos categóricos [5]. Esto permite que la información fonética sea clasificada de una forma más eficiente, optimizando los resultados obtenidos. Otra ventaja del planteo utilizado, es que al utilizar GMM se asume que los datos a procesar tienen una distribución gausiana, lo que no es válido, ya que como puede apreciarse en [8], dicha información no tiene ninguna distribución específica. Por otro lado, el sistema KH, trabaja en base a K-Histogramas y, por ende, se ajusta a la distribución probabilística de los datos a procesar.

Este sistema nuevamente posee dos etapas, una de entrenamiento y otra de conversión. En la primera se particiona el espacio de cada uno de los locutores con el fin de generar los K-Histogramas que se utilizarán durante la conversión. Luego, en la etapa de entrenamiento, se toman los segmentos de voz a convertir y mediante una transformación como la que se aprecia en la Figura 5 se obtiene una estimación *y' = F(x)*, la cual se utilizará para el sintetizado de la voz convertida.

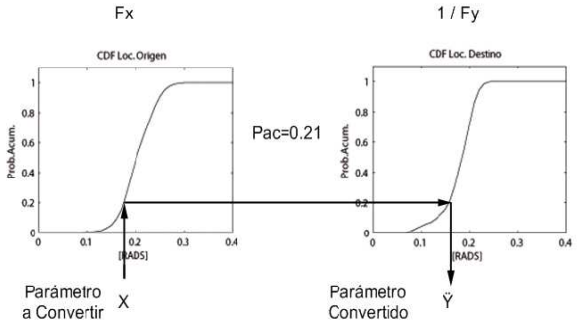

Fig. 5. Transformación Propuesta.

Un aspecto que debe tenerse en cuenta es que a la hora de obtener los parámetros a utilizar para realizar la transformación, debe realizarse una discretización del espacio. Por este motivo, las funciones acumuladas de probabilidad correspondientes a cada uno de los locutores son discretizadas. En este trabajo se realiza dicha discretización utilizando 314 y 3140 divisiones o *bins*. Esto genera las siguientes consecuencias: Discretizar utilizando mas divisiones, el ancho de las mismas se reduce y por lo tanto el error debido a la cuantificación disminuye. Si, por otro lado, la discretización se realiza utilizando menos divisiones, si bien aparece un cierto error debido a la cuantificación, el costo computacional se reduce notablemente.

En la Figura 6 se aprecia un diagrama en bloques del sistema propuesto.

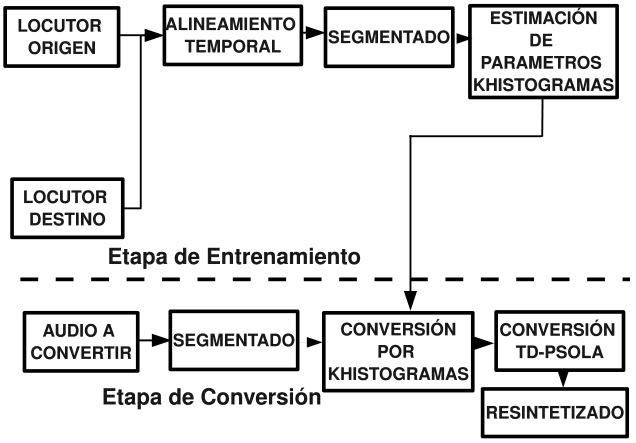

Fig. 6. Sistema de VC basado en K-Histogramas.

Este tipo de sistemas permite obtener voces sintetizadas con una identidad mas elevada respecto a las obtenidas utilizando GMM y FS. Dado el propósito de este trabajo, se estudian otros métodos con el fin de optimizar el rendimiento del sistema. Es por ello, que a lo largo de las siguientes subsecciones se estudian dos métodos que combinan partición del espacio con selección de unidades.

C. Sistemas de Conversión de Voz que combinan Selección de Unidades y partición del Espacio.

Con el fin de aumentar la relación identidad-calidad, se diseñaron sistemas que combinan ambas técnicas previamente descriptas. De este modo, estos sistemas combinan una primera etapa que utiliza partición del espacio con una segunda etapa que opera mediante selección de unidades. A continuación se presentan dos sistemas que operan en base a esta filosofía.

En la Figura 7 se presenta el sistema propuesto por Dutoit [7], el cual se caracteriza por estar compuesto por dos etapas de conversión, una compuesta por un bloque de partición del espacio mediante GMM y una segunda conversión mediante selección de unidades.

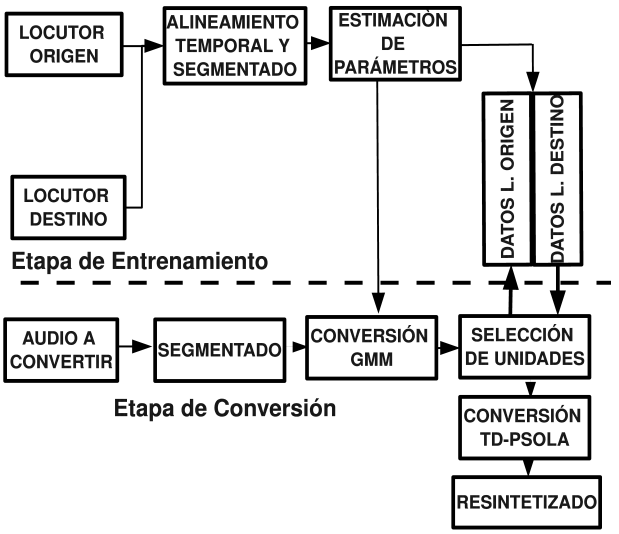

Fig. 7. Sistema de Conversión de Voz propuesto Dutoit.

Los experimentos realizados con este sistema, permiten obtener voces sintetizadas que mejoran los niveles de identidad de los sistemas GMM, pero con un deterioro considerable de la calidad de los audios sintetizados.

Por otro lado, el algoritmo de conversión de voz KHFS [8], presenta un enfoque similar al propuesto por Dutoit. En este caso, el bloque de partición del espacio opera utilizando el algoritmo K-Histogramas. Nuevamente, este es seguido por un bloque de selección de unidades. Una representación de este tipo de sistemas se presenta en la Figura 8.

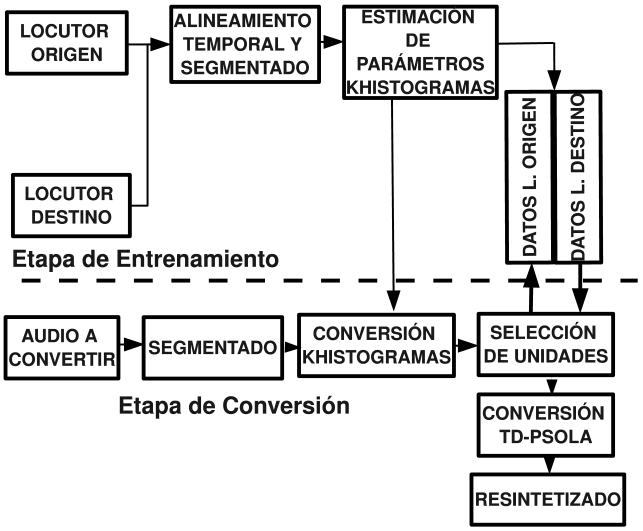

Fig. 8. Sistema de Conversión de Voz KHFS.

Los resultados obtenidos por este método en cuanto a la identidad de las voces convertidas debería ser el más elevado de los estudiados.

Sistemas como el propuesto por Dutoit [7] o el desarrollado en [8], presentan el problema de requerir tiempos de procesamiento elevado, debido al funcionamiento del bloque de selección de unidades.

Debido a que se debe seleccionar un sistema de modo tal que la identidad de la voz convertida tenga una similaridad mas elevada respecto a la del locutor destino que se desea obtener (en este caso, la voz a obtener es la original de la persona afectada por la dolencia). Es por ello, que en la siguiente Sección se realizan experimentos comparativos entre los sistemas de conversión de voz presentados con el fin de seleccionar el más apropiado para la implementación del sistema.

IV. Resultados Experimentales

Con el fin de seleccionar el sistema de conversión de voz mas apropiado para la implementación del sistema, se realizan dos experimentos subjetivos. Para ello se utilizaron 205 oraciones en español con una duración promedio de 4 segundos, cada una de las cuales fue pronunciada por cuatro locutores, dos hombres y dos mujeres. De estas 205 oraciones, 150 fueron utilizadas para el entrenamiento y 55 para la evaluación de los sistemas,

Las oraciones tienen fs=16000Hz y fueron procesadas utilizando las técnicas previamente descriptas. Los audios sintetizados fueron analizados por un grupo de 15 evaluadores voluntarios (con consentimiento informado), quienes los evaluaron en **Identidad** y **Calidad** con puntajes del 1 (identidad o calidad pésima) a 5 (identidad o calidad

excelente). A partir de ello se genero un Test MOS (*Mean Opinion Score*) [12] para ambas características: Calidad (MOS-Q) e Identidad (MOS-S).

En la Tabla 1 se aprecia, en la primera columna los sistemas evaluados. En la segunda columna **MOS-S**, es decir el test MOS de identidad, mientras que en la tercera columna se aprecia **MOS-Q**, el test MOS de Calidad.

Sistema	MOS-S	MOS-Q
KH3140	3.6	3.0
KH314	3.6	2.8
FSKH3140	3.1	2.3
FSKH314	3.4	2.3
DUTOIT	3.4	2.4
GMM	2.7	2.1
FS	3.6	1.8

Tabla 1. Resultados Experimentales.

Los resultados muestran una clara ventaja de los métodos basados en K-Histogramas, un detalle relevante es que el exceso de manipulación debido al bloque de selección de unidades en los sistemas FSKH deteriora sus resultados tanto de inteligibilidad como de calidad. Otro aspecto que debe ser tenido en cuenta a la hora de implementar el sistema es que si bien el método KH con 3140 *bins* es el que obtiene el mejor resultado, éste tiene una carga computacional casi diez veces mayor que el sistema KH con 314 *bins*, lo que incrementa notablemente el tiempo de procesamiento.

Por otro lado, se realizaron test de Wilcoxon para los sistemas bajo estudio, en estos se verificó que en los experimentos MOS-S y MOS-Q, los sistemas **KH314** y **KH3140** son similares al modelo ideal con $\boldsymbol{p < 0{,}01}$.

Debido a los resultados obtenidos en los experimentos, se selecciona el sistema **KH314** para implementar el sistema de asistencia oratoria.

V. Implementación del Sistema

El sistema implementado puede apreciarse en el sitio de internet http://200.0.183.36:8080 . En dicha plataforma el usuario puede escribir un texto y sintetizarlo utilizando dos voces: una obtenida mediante el sistema TTS FESTIVAL [9] o una alternativa, generada mediante conversión de voz utilizando la técnica basada en K-Histogramas **KH314**. Un esquema simplificado del sistema se presenta en la Figura 9.

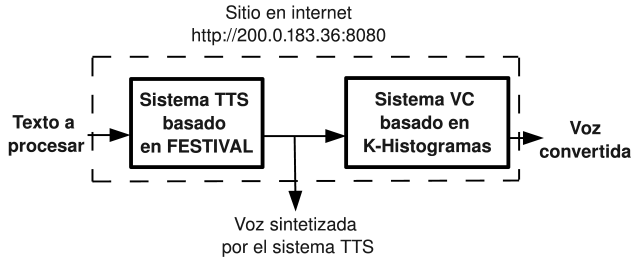

Fig. 9. Sistema Implementado.

VI. Conclusiones

A lo largo del presente trabajo se desarrolló un sistema capaz de transformar un texto ingresado por el usuario en la voz de una persona en particular la aplicación FESTIVAL. Luego, mediante técnicas de conversión de voz basadas en K-Histogramas, fue posible transformar dicha voz en otra, como por ejemplo la de un paciente con una discapacidad fonatoria.

Se pudo comprobar que la técnica de K-Histogramas con 314 *bins* es la mas adecuada por su relación Identidad de voz – Calidad - Tiempo de procesamiento.

En el futuro se pretende optimizar el sistema con el propósito de obtener voces con una identidad más elevada, en un tiempo de procesamiento que se acerque aún más a tiempo real.

Agradecimientos

Queremos agradecer a los miembros del Laboratorio de Comunicaciones (LAC) que colaboran con nuestro trabajo.

Referencias

[1] M. Abe, S. Nakamura, K. Shikano, "Voice Conversion through vector quantization", *En Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 655-658, 1988.

[2] M. Abe, "A Segment-based approach of Voice Conversion", *En Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 765-768, 1991.

[3] H. Duxans i Barrobés, "Voice conversion applied to text-to-speech synthesis". *En PhD Thesis (Universitat Politècnica de Catalunya)*, 2006.

[4] Y. Stylianou, O. Cappe y E. Moulines, "Continuous probabilistic transform for voice conversion", *En Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol 6 no 2, pp 131-142, 1998.

[5] A. A. Uriz, P. Agüero, J.C. Tulli, E.- González y A Bonafonte I Càvez, "A Comparison Between GMM and non-GMM models applied in Voice Conversion", *En Proceeding of 10º Argentine Symposium on Technology*, pp 31-44, 2009.

[6] D. Erro y A. Moreno, "Weighted Frequency Warping for Voice Conversion", *In Proceedings of theInterspeech 2007,* 2007.

[7] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet,J. Perez y Y. Stylianou, "Towards a voice conversionsystem based on frame selection". *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp 513-516, 2007.

[8] A. Uriz, P. Agüero, J.C. Tulli y A Bonafonte I Càvez, "Voice Conversion using K-Histograms and Frame Selection", *En Proceeding of Interspeech 2009*, 2009.

[9] A. W. Black y K. A. Lenzo. "Building Synthetic Voices". 1999.

[10] J.A. Gurlekian, L. Colantoni y H. Torres. "El alfabeto fonético SAMPA y el diseño de corpora fonéticamente balanceados". *Fonoaudiológica. Editorial: ASALFA. Tomo: 47, Numero: 3, pp 58-69,* Diciembre 2001.

[11] S. P. Taylor, P. Taylor, A. W. Black y R. Caley. "The Architecture Of The Festival Speech Synthesis". *In The Third ESCA Workshop in Speech Synthesis. pp. 147—151.* 1998.

[12] International Telecomunication Union. "Series P: Telephone Quality, Telephone Installations, Local Line Networks: Methods for objective and subjective assessment of quality*". Ed. International Telecomunication Union. Ginebra, Suiza.* 2003.