International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org Vol. 10 No. 1 August 2018, pp. 209-217

Optimización de Síntesis de Voz mediante Análisis de Componentes Principales

[Speech Synthesis Optimization by Principal Component Analysis]

Manuel Haro Marquez

Laboratorio de software libre
Av. de la Juventud 504, Barros Sierra, La Encantada, 98090, Zona A,
La Encantada, 98090 Zacatecas, Zac.

Miguel Ángel Fraire Hernández,Luis Alberto Flores Chaires, María del Rosario Bernal Díaz

IGA, Instituto de Graduados en Administración
Avenida Universidad 137, Lomas del Patrocinio, 98060 Zacatecas, Zac.

Fabian Navarrete Rocha, Noe Buenrostro Domínguez, Ana Lourdes Borrego Elías, Ana Luisa Hernández Gutiérrez, Nadia Garibay Rendón, Lorena Raquel Casanova Luna, Jaquelinne Martínez Robles, Luis Enrique Paredes Berumen, Jose Manuel Cervantes Miramontes

Centro de Investigación e Innovación Automotriz de México (CIIAM)
Ramón López Velarde 801, Zacatecas Centro, 98000 Zacatecas, Zac.

Claudia Sifuentes Gallardo

Centro de Investigaciones en Óptica (CIO)
Lomas del Bosque 115, Lomas del Campestre, 37150 León, Gto.

Margarita Selene Salazar Ávila

Centro de Investigación para la Administración Educativa
Independencia 1100, Centro, 78000 San Luis, S.L.P.
Av. Cerro del Padre 907, Zacatecas, Zac.

Abubeker Gamboa Rosales

Universidad Tecnológica del Estado de Zacatecas (UTZAC)
Km 5, Carretera Zacatecas, 98601 Cuauhtémoc, Zac.

César Gamboa Rosales

Universidad Aeronáutica en Querétaro
Carr. Querétaro-Tequisquiapan, Coyote, Qro.

***Resumen* — Los sistemas de síntesis de voz por concatenación proporcionan una alta calidad debido a la mejora en sus métodos de selección de unidades acústicas. Sin embargo, la calidad de síntesis de voz depende de la posición adecuada de los puntos de concatenación en la secuencia de unidades acústicas a concatenar que componen una señal de voz sintetizada. La posición de los puntos de concatenación en gran medida determina el grado de distorsión percibida por el humano en una voz sintética. Por lo tanto, presentamos un algoritmo de fijación del punto óptimo de concatenación entre unidades acústicas que se basa en el análisis de componentes principales para la reducción de la distorsión acústica percibida por el oyente en un sistema de síntesis de voz. Este**

Corresponding Author: Fabian Navarrete Rocha

algoritmo propone la extracción de parámetros espectrales en referencia a un punto de concatenación y su transformación mediante el análisis de componentes principales en una representación parsimoniosa la cual representa las propiedades más relevantes de la señal de voz. Posteriormente, se determina el punto óptimo de concatenación mediante un proceso iterativo de mejora continua. Finalmente, los resultados experimentales del método de punto de concatenación optimización propuesto son analizados.

Palabras clave **- Síntesis de voz, selección de unidades, concatenación, percepción de discontinuidad.**

I. INTRODUCCIÓN

El método de síntesis voz basado en la concatenación de unidades ha sido utilizado por los sistemas electrónicos conocidos como Texto-a-Voz (TTS) en los últimos años [1]. En los cuales el método de síntesis concatenativa se basa en un inventario acústico de unidades de voz. Dichas unidades deben de cubrir una gran variedad de características fonéticas y prosódicas del lenguaje a sintetizar; permitiendo al selector de unidades de voz del sintetizador extraer la más adecuada secuencia de unidades de voz contenidas en el inventario para sintetizar un texto de entrada; y de esta forma minimizar las distorsiones entre las unidades a concatenar. Las unidades de voz buscadas por el selector de unidades por lo regular son extraídas de diferentes contextos fonéticos y prosódicos dentro del inventario, lo cual provoca que se presenten discontinuidades y desfases en su forma espectral dentro de los bordes o fronteras donde son concatenadas. Adicionalmente, las unidades de voz extraídas tienen una longitud variable al ser fonemas, difonemas, trifonemas, etc.; lo que origina un número no determinado de puntos de concatenación a optimizar dentro de una secuencia de texto. Debido a que las discontinuidades entre las unidades de voz concatenadas deterioran significativamente la calidad de síntesis del habla, la investigación y desarrollo de un método de concatenación óptimo se ha convertido en una tarea apremiante en la síntesis de voz [2]. La percepción humana de distorsión en una señal sintetizada se estima mediante el proceso de selección de unidades a través del cálculo del costo de concatenación. Dicho costo es la suma ponderada de n subcostos de concatenación tales como la diferencia espectral basado en la transformación rápida de Fourier (FFT), coeficientes de frecuencias espectrales lineales (LSF) y Mel-coeficientes de frecuencia cepstral (MFCC). Sin embargo, los subcostos de concatenación antes mencionados derivan del análisis de Fourier y cada uno está relacionado con el nivel de percepción humana de distorsión en una señal sintética a través de la estimación de la discontinuidad espectral entre los puntos fijos de concatenación de una secuencia de unidades de voz. Por lo anterior, la apropiada determinación de los puntos de concatenación entre las unidades a concatenar evita la aparición de artefactos y distorsiones en la voz sintetizada obteniendo como resultado una mayor calidad de síntesis de voz. Por lo anterior, se propone un algoritmo para determinar el punto óptimo de concatenación entre un conjunto de unidades de voz dentro de un inventario basado en el análisis de componentes principales (PCA); y de esta forma disminuir eficazmente que emerjan artefactos o distorsiones en una señal de onda sintetizada debido a concatenaciones no apropiadas. PCA representa una alternativa de extracción de información de las unidades de voz y proporciona un análisis innovador para la caracterización de discontinuidades acústicas entre unidades acústicas. PCA es utilizado para extraer información característica dentro de las fronteras de concatenación de las unidades de voz y es representada a través de una estructura predeterminada. Esta estructura es utilizada por el proceso de optimización del punto de concatenación (OPC) para determinar los puntos óptimos de concatenación de cada unidad en referencia a todas las demás unidades de voz en el inventario. El OPC se realiza offline y coadyuva en el proceso del selector de unidades al incrementar la calidad de síntesis de voz. La Sección 2 muestra una descripción amplia de un sintetizador de voz. Posteriormente en la Sec. 3 se analiza la descomposición modal PCA de la información y las principales características a utilizar en el proceso OPC. En seguida la Sec. 4 ofrece una visión descriptiva del proceso OPC. Finalmente, las Sec. 5 y 6 reportan el análisis de los resultados experimentales del algoritmo OPC entre unidades de voz y conclusiones respectivamente.

II. SISTEMA DE SÍNTESIS DE VOZ

En esta sección se describe en forma esquemática las componentes de un sistema de síntesis de voz mostradas en la Fig. 1. Esta muestra las principales componentes del bloque de procesamiento de texto a nivel simbólico dentro de un sistema de síntesis de voz. Al inicio del diagrama, el texto de entrada es procesado en diferentes secciones. Plain ASCII-text o texto enriquecido con información conceptual es la componente que contiene las formas de pronunciación y acentuación especiales de palabras correspondientes a un lenguaje. Dentro de este, el texto de entrada es procesado por la etapa de pre-procesamiento del sistema. De la misma forma, las fronteras de palabras, frases y oraciones son clasificadas y etiquetadas.

Adicionalmente, las combinaciones de caracteres especiales son reconocidas tanto como las palabras de función y abreviaturas son detectadas e identificadas dentro del texto en tiempo de ejecución. Las reglas para el procesamiento de formatos numéricos y anomalías de texto como abreviaturas y nombres son analizadas inicialmente, y así se genera una secuencia fonológica del texto de entrada. Posteriormente, el resto del texto es tokenizado y por medio de un léxico especial los tokens son procesados. Si el proceso aún no se ha realizado correctamente, el léxico de abreviaturas es aplicado para cada token.

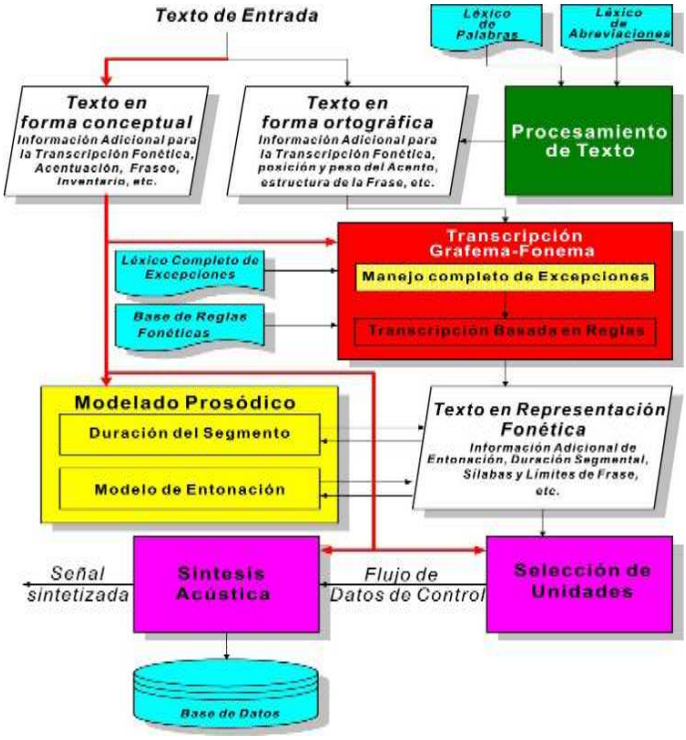

Fig. 1 Sistema de Síntesis de Voz

Finalmente, la detección del límite de frase proporciona una subdivisión más precisa de las oraciones dentro de frases prosódicas. La etapa de conversión grafema-fonema deriva en una representación fonética del texto de entrada. El procesamiento se realiza primero por un léxico base y luego por una componente basada en reglas de conversión. La construcción de las reglas fonéticas sigue un enfoque descrito en [3].Todas las reglas fonéticas se organizan en prefijos grafémicos, cuerpo de reglas de conversión, sufijos y en resultados de fonetización. Adicionalmente el lugar y tipo de acentos se suministran mediante los siguientes componentes prosódicos.

A. Modelado Prosódico

Procesamiento prosódico (controles de duración y entonación) se realiza mediante varios módulos que agregan la duración y parámetros de entonación en la cadena de información fonética por segmento. El modelo de entonación se realiza mediante una red neuronal (ANN) de datos, la cual es utilizada para generar estilos flexibles de habla y su vez adaptar rápidamente el sistema de síntesis de voz a los requerimientos de diferentes tipos de voces e idiomas. Por medio de la red neuronal se estima directamente el contorno de la frecuencia fundamental f0 de una secuencia de vectores lingüísticos de entrada. La codificación de información prosódica está orientada silábicamente, desde la secuencia de fonemas, y las silabas son aisladas e introducidas iterativamente a una red concurrente (incluyendo la silaba focal y dos silabas correspondientes al contexto previo y el posterior, lo que significa un marco de contexto de C=5), [4].Para cada silaba, un vector con características lingüísticas y fonéticas N1=8 se aplica en la red de entrada. La primera capa oculta consiste en N2=10, la segunda capa oculta de neuronas N3=6. La segunda capa oculta está conectada por completo a las neuronas de contexto, es decir, la capa de entrada ANN contiene C*N1+N3=46 neuronas. La capa de salida posee N4=3 neuronas, que estiman el contorno de la frecuencia fundamental de la sílaba focal. El codificador de entrada considera la posición en la frase, estrés, características fonéticas del núcleo y contexto como se muestra en la Fig. 2.

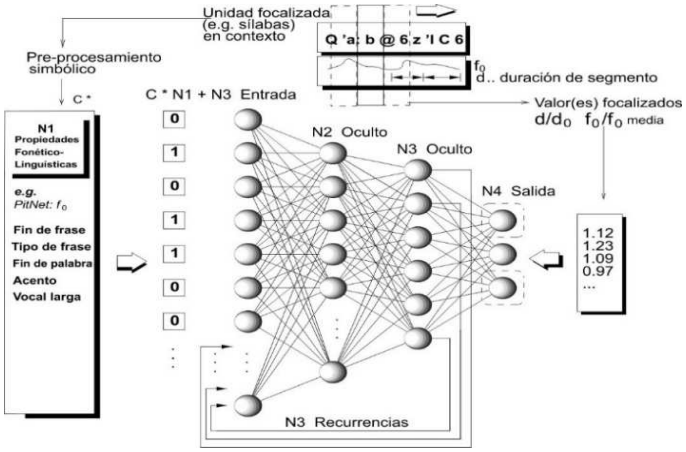

Fig. 2 ANN del modelo de control de entonación.

B. Selección De Unidades

El módulo de selección de unidades tiene como función principal el encontrar dentro de la base de datos la combinación más adecuada de secuencias de unidades de voz que aseguren que las diferencias de percepción humana entre

la esperada señal de voz natural y la señal de voz sintetizada sean lo más mínimas. Este módulo utiliza la información de la secuencia de fonemas procedentes del modelado prosódico para realizar la búsqueda de las secuencias más apropiadas; y una vez seleccionada las unidades de voz; a las cuales se les agrega la información prosódica correspondiente. En el método de selección de unidades se utilizó el selector de unidades propuesto en [5].

C. Síntesis Acústica

Finalmente, la síntesis acústica construye una señal de voz sintética a partir de una secuencia de unidades seleccionadas y a su vez la modifica y reproduce conforme a los contornos de parámetros prosódicos modelados. Dentro del módulo de síntesis acústica se pueden realizar algunas manipulaciones prosódicas en lo referente al contorno de la frecuencia fundamental, duración y energía de los fonemas que conforman la señal de voz sintetizada. Sin embargo, ambos mecanismos de modificación de frecuencia fundamental, así como duración de un fonema se comportan de forma contraria. La duración de un fonema es la suma de la duración de sus períodos que obviamente están cambiando mientras se manipula la frecuencia fundamental. Esto implica que su frecuencia fundamental no se puede modificar de forma independiente a la duración. Es por ello que la modificación prosódica de señales de voz se da mediante un proceso iterativo hacia los valores objetivos [6].

III. Estructura Basada En El Análisis PCA Voz

Las unidades de voz básicas en la síntesis de voz son los difonemas; y han sido utilizados en el desarrollo de sintetizadores de voz. Esto se debe a que al realizarse la concatenación en la parte media del fonema que conforma el difonema, esta tiende a ser más suave. El primer paso para la conformación de la estructura representativa de las unidades de voz es la agrupación de los fonemas que se encuentran en los enunciados grabados en el inventario (por ejemplo, /O/). De esta forma es posible abstraer los puntos de concatenación óptimos dentro de las unidades de voz y reducir el desfase espectral y las posibles distorsiones cuando estas son concatenadas. La Fig. 3 muestra dos unidades de voz denotadas por L1-R1 y L2-R2 donde L1 representa la primera mitad y R1 la segunda mitad contigua del fonema /O/1, y L2 representa la primera mitad y R2 representa la segunda mitad contigua del fonema /O/2, [7].

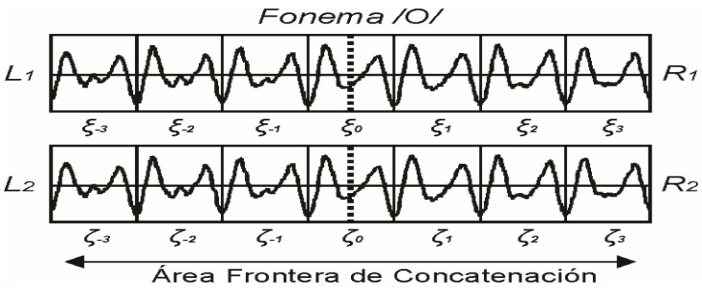

Fig. 3 Marcado de los Periodos Fundamentales Centrados de Tono (Periodos de Pitch Centrados) Notación K=3.

El objetivo es encontrar el punto óptimo de concatenación entre L1 y R2, que al concatenarlas formaran una unidad acústica de voz (difonema) libre de distorsiones o artefactos. Sean ξ-K+1,…, ξ0,…, ξK-1 y z-K+1,…,z0,…,zK-1 los 2K-1 periodos fundamentales centrados de Tono (Pitch) asociados con el área de concatenación de las unidades de voz L1-R1 y L2-R2, respectivamente. Aunado a ello, los limites o fronteras entre las unidades L1-R1 y L2-R2 se sitúan exactamente en medio de ξ0 y z0 como se muestra en la Fig. 1.La extracción y marcado de los periodos fundamentales es obtenida para unidades fónicas y afónicas a través del algoritmo propuesto en [8]. Para ejemplificar la conformación de la estructura representativa de las unidades de voz del inventario que posteriormente será utilizada por PCA; se considera que el inventario contiene M1 unidades de voz, denotadas como L1-R1 y M2 unidades de voz, denotadas como L2-R2, con un punto de concatenación dentro del fonema /O/ dentro del completo inventario. Esto implica la búsqueda de un punto óptimo de concatenación a través de (2K-1)(M1×M2) periodos fundamentales centrados asumiendo M1×M2=M combinaciones de concatenaciones y N como el número máximo posible de muestras entre los periodos fundamentales por unidad de voz [9]. Adicionalmente, rellenado con ceros simétrico (Zero-padding) es aplicado, en caso de ser necesario, así como ventaneo ponderado para todos los periodos fundamentales centrados N.La matriz resultante es X((2K 1)M×N) con elementos xi,j, donde xi representa los periodos fundamentales centrados y la columna xj representa el segmento de muestras en el dominio del tiempo como lo se expone en [10]. Posteriormente, se transforma la matriz de datos de entrada X por medio del análisis PCA, El análisis PCA descompone el conjunto de datos de señales mixtas de entrada en un conjunto de datos de señales no correlacionadas. En términos de momentos estadísticos, PCA estima una matriz que representa la transformación de los datos de entrada X^((x_1^1,x_2^2,…,x_(2K-1)M^N)) con una función de densidad de probabilidad (pdf) denotados por p^((x_1^1,x_2^2,…,x_(2K-1)M^N)) en un conjunto de

señales no correlacionadas Y^((x_1^1,x_2^2,…,x_(2K-1)M^N)) como se muestra a continuación en la ecuación (1).

$$Y = A\,(X - \mu_x) \tag{1}$$

Donde μ_x representa la media de la matriz de entrada X y la matriz A se compone de los vectores propios de la matriz de covarianza de la matriz de entrada X tal y como se ilustra en las filas de la matriz transpuesta A^T en la Fig. 4

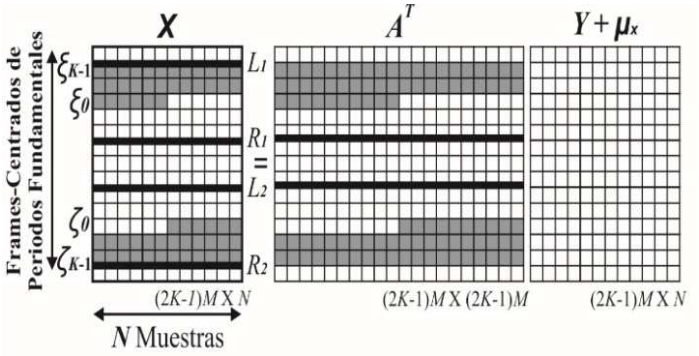

Fig. 4 Representación de la matriz de entrada X

La estimación de la media μ_x, es parte fundamental de la solución en la búsqueda de la matriz base de componentes principales, la cual permite minimizar el error cuadrático medio en una aproximación de datos. Posteriormente, es posible reconstruir la matriz de entrada X mediante el uso de la propiedad ortogonal $A^T=A^{-1}$ y la matriz Y como se ilustra en la siguiente ecuación (2).

$$X = A^T(Y + \mu_x) \tag{2}$$

Donde A es la base ortogonal de tamaño (2K-1)M×(2K-1)M para el análisis PCA en el espacio L comprendido por (2K-1)M-dimensional. La matriz Y puede ser vista como la matriz de coordenadas de la base ortogonal. La matriz de entrada X es proyectada sobre los ejes de coordenadas definidos por la base ortogonal. De esta forma la matriz de entrada X (la cual contiene todas las unidades de voz candidatas) está representada por una combinación lineal de los vectores de la base ortogonal en el espacio L con dimensión M [11].

Esto puede ser descrito como una descomposición de rango-M, que define una asignación o mapeo entre el conjunto de los periodos centrados de tono y el conjunto de (2K-1)M-dimensional vectores característicos $\tilde{a}_i = a_i$ (Y+M_X)).. La estructura propuesta ofrece una metodología de extracción de información característica con una visión global sobre todas las unidades de voz del inventario acústico donde los componentes de los vectores propios interactúan directamente en los límites o fronteras de concatenación en el fonema /O/ tal como se ilustra en la Fig. 2. Basados en la estructura propuesta, el proceso de optimización OPC de las unidades de voz candidatas representadas a través de vectores característicos se determinará por las características generales estimadas en los periodos fundamentales en un marco analítico basado en PCA, y no estará restringido a una instancia específica como lo es análisis característico en dominio de la frecuencia. Por lo tanto, la unión de dos vectores aς y aξ en un espacio característico nuevo (transformación PCA) refleja un alto grado de similitud, y por lo tanto potencializa la estimación un grado de desajuste o distorsión en la concatenación de unidades de voz.

A. Análisis Del Punto De Concatenación

El proceso de búsqueda de los puntos óptimos de concatenación de las unidades de voz en un inventario debe llevarse a cabo sobre todos los M posibles límites de concatenación de los segmentos L_1-R_1, L_2-R_2, L_3-R_3,…,L_M-R_M de las unidades de voz candidatas. OPC debe buscar los puntos de concatenación óptimos basándose en las aéreas fronteras a concatenar entre todas las combinaciones posibles L_1-R_2; y de esta forma establecer M concatenaciones con mínima distorsión cuando sean llevadas a cabo. Posteriormente, las unidades de voz que presentan la mínima distorsión en su punto de concatenación optimizado son extraídas por el selector de unidades de voz. Para llevar a cabo esta tarea los puntos candidatos de concatenación de las unidades son representados en el espacio L, [12], [13], [14].

La transformación PCA comprende los vectores ⟦a_x⟧_k y ⟦a_z⟧_k, los cuales representan los periodos fundamentales centrados de tono xk y zk para -K+1≤ k ≤ K-1. Considerando ahora el punto candidato de concatenación entre L1-R2 como se muestra en el área sombreada en la Fig. 2. El cual es descrito tal x-K+1,…,x1, φ0,z1,…, zK-1 donde φ0 ilustra el periodo centrado concatenado. Este se compone de la mitad de la parte izquierda de x0 y de la mitad de la parte derecha de z0; y se representa en el espacio L mediante la transformación lineal PCA de la matriz de entrada X como se muestra a continuación.

$$\tilde{\mathbf{a}}_{\xi_{-K+1}}, \ldots, \tilde{\mathbf{a}}_{\xi_1}, \tilde{\mathbf{a}}_{\varphi_0}, \tilde{\mathbf{a}}_{\zeta_1}, \ldots, \tilde{\mathbf{a}}_{\zeta_{K-1}} \tag{3}$$

Donde ãx representa la parte izquierda de la unidad L1 en la matriz A y ãz representa el lado derecho de la unidad R2 en la matriz A. Sin embargo, $\tilde{a}(\varphi\_0)$ no tiene una representación en la matriz de entrada X, pero puede ser calculada analizando

φ0 como un renglón de la matriz característica de la matriz de entrada X como se muestra en la ecuación (4).

$$\varphi_0 = a_{\varphi_0} R = \tilde{a}_{\varphi_0} \tag{4}$$

Donde el vector (2K-1)M-dimensional a_(φ_0) es introducido como un renglón adicional en la matriz A, y R es el operador correspondiente de Y+mx. Así, el vector de concatenación a ̃_(φ_0)=a_(φ_0) R representa φ0 en el espacio L. Una vez que a ̃_(φ_0)es el estimado del grado de desajuste o distorsión en la concatenación de unidades de voz, este se calcula como la diferencia acumulada entre los vectores compuestos de dos segmentos de unidades de voz a ̃_(x_1)y a ̃_(z_1). Y se alcanza mediante la medición de similitud entre dos vectores individuales como se propone:

$$s(\tilde{a}_k, \tilde{a}_1) = \cos(a_k R, a_1 R) = \frac{a_k R R^T a_1^T}{\|a_k R\| \|a_1 R\|} \tag{5}$$

Donde k≥1 y l≤(2k-1)M. Si utilizamos la notación abreviada a partir de (Galvan, 2016) se obtiene como resultado:

$$s(a_{\propto -K}, a_{\propto 0}, a_{\propto K}) = \frac{s(a_{\propto -K}, \tilde{a}_{\propto 0}) + s(\tilde{a}_{\propto 0}, a_{\propto K})}{2} \tag{6}$$

Finalmente, el promedio de similitud entre los vectores frontera de las dos unidades concatenadas puede ser descrito de la siguiente forma:

$$d(V_1,V_2) = \sum_{k=1}^{K-1} 2\,\tilde{s}\left(a_{\xi_k}, a_{\xi_\varphi}, a_{\zeta_k}\right) - \tilde{s}\left(a_{\xi_k}, a_{\xi_0}, a_{\xi_{-k}}\right) - \tilde{s}\left(a_{\zeta_{-k}}, a_{\zeta_0}, a_{\zeta_k}\right) \tag{7}$$

Esta estimación de similitud corresponde a la medida de la diferencia de trayectoria antes y después de la concatenación a través de toda el área límite de concatenación. Esta medida puede ser probada cuando la medida de diferencia entre dos unidades de voz contiguas en la base de datos es calculada, por ejemplo: x´s es igual a z´s. Lo que significa que φ0=x0=z0 y d(V1,V2)=0 si, y solo si, L1=R2. Lo que significa que entre más cerca este la estimación de similitud a cero, el punto de concatenación será más adecuado entre dos unidades a concatenar [15], [16], [17].

IV. Optimización Del Punto De Concatenación

Cuando la estructura analítica PCA para el OPC es establecida, el siguiente paso es desarrollar el procedimiento para determinar los puntos de concatenación óptimos y las unidades de voz que presentan la menor distorsión cuando éstas sean concatenadas [18], [19]. La propuesta consiste en calcular la distancia acumulada d(V1,V2) centrándose en cada punto de concatenación posible en el área [-K,K], utilizando el espacio L asociado al área de concatenación ilustrado en el diagrama de flujo de la Fig. 5.

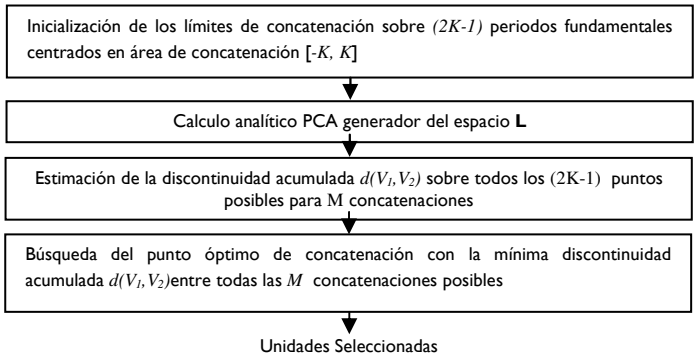

Fig. 5 Optimización del Punto de Concatenación (OPC).

Primeramente, la etapa de inicialización establece el marco de información de periodos fundamentales centrados 2K-1 sobre el área de concatenación [-K, K]. Después se deriva el marco de información de datos de periodos fundamentales centrados de Tono dentro el espacio L utilizando PCA. El resultado provee (2K-1)M vectores característicos sobre el espacio L, con el mismo número de puntos potenciales de concatenación. Posteriormente, se calcula la discontinuidad acumulada asociada a cada punto de concatenación para todos los puntos de concatenación candidatos M. Finalmente, el punto de concatenación y las unidades que presentan la discontinuidad mínima acumulada son seleccionadas por el selector de unidades de voz [20], [21]. Ambos procedimientos como la estructura PCA y OPC son descritos a detalle en la Fig. 6.

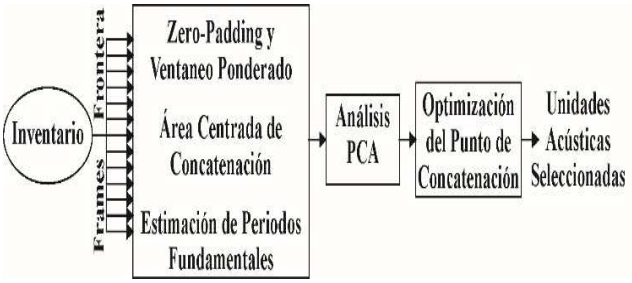

Fig. 6 Estructura PCA y Procedimiento OPC.

Donde las estructuras (frames) de datos de periodos fundamentales centrados son extraídas de las áreas fronteras de concatenación de las unidades de voz y son representadas en el espacio L utilizando análisis PCA. Finalmente, los

puntos de concatenación con la menor discontinuidad acumulada d(V1,V2) son seleccionados entre todos los puntos de concatenación posibles, lo que representa el punto de concatenación óptimo de cada correspondiente unidad de voz dentro del inventario y es utilizado en el proceso de selección de unidades de un sintetizador de voz TTS.

V. Resultados Experimentales

En la síntesis de voz concatenativa se observa una mayor tasa de detección de discontinuidades en los diptongos vocales que en los monoptongos vocales.En consecuencia, nos centramos en el análisis de la distorsión de concatenación del método propuesto OPC en la síntesis de los diptongos ingleses /eI/, /OU/, /aI/, /aU/ y /OI/. Para ello, utilizamos el Sistema Dress-TTS y la base de datos del habla inglesa "TC-STAR" y de esta forma sintetizar los diptongos antes mencionados. La síntesis de diptongos fue realizada por medio de la unión de dos difonemas los cuales al concatenarlos forman el diptongo deseado y a su vez estaban incluidas, en una palabra. Para realizarlo, en primer lugar, se extrajeron todas las unidades de voz en el inventario con la frontera izquierda o derecha situada en la mitad del fonema correspondiente a un diptongo. Fueron en total 2902casos para /eI/, 2094 para /OU/, 3373 para /aI/, 1496 para /aU/ y 644 para /OI/.Adicionalmente, se extrajeron K=5 periodos de Tono sobre el lado izquierdo y derecho de las áreas frontera de concatenación, resultando 2K-1=9 periodos centrados de Tono. Posteriormente, se siguió el procedimiento descrito en la sección anterior para determinar el punto óptimo de concatenación sintetizando el diptongo /eI/ dentro de la palabra “same”. Con esto se obtuvo una matriz de datos de entrada X de tamaño (645×128); la cual se compuso de 28 y 47 unidades de voz para el lado izquierdo y derecho respectivamente; dentro de todas las posibles combinaciones de concatenación. Enseguida se aplicó el proceso OPC para el diptongo /OU/ dentro de la palabra “hope”.Así, se conformó la matriz de entrada de datos X con un tamaño de (1199×128) con 36 y 96 unidades de voz para el lado izquierdo y derecho respectivamente. El proceso OPC para los diptongos /aI/ dentro de la palabra “Kite” proporciono una matriz de entrada de datos X(3553×128) con 45 y 372 unidades de voz para las fronteras izquierda y derecha. De la misma manera OPC fue aplicado para el diptongo /aU/ dentro de la palabra “house” proporcionando una matriz de entrada de datos X(1296×128), la cual es integrada para 83 y 61 unidades de voz para el lado izquierdo y derecho respectivamente dentro de las combinaciones posibles de concatenación. Finalmente, se obtuvo la matriz de entrada de datos X para el diptongo /OI/ dentro de la palabra “join” con un tamaño de (144×128), la cual integra 5 y 11 unidades de voz. La Fig. 7 muestra el diptongo sintetizado /OI/ dentro de la palabra “join” utilizando el proceso OPC.

La Fig. 7 está compuesta de la representación de la señal de voz del diptongo /OI/ de la palabra “join” en el dominio del tiempo, así como su representación en un espectrograma y sus frecuencias espectrales lineales en el dominio de la frecuencia. En el espectrograma no se observa discontinuidad entre sus frecuencias espectrales líneas y ninguna discontinuidad relevante en el formante espectral correspondiente al área de concatenación del diptongo /OI/ sintetizado. Adicionalmente, ninguna distorsión de concatenación es percibida al escuchar esta palabra. El resultado del proceso OPC para la palabra “house” que contiene el diptongo /aU/ es mostrado en la Fig. 8. De la misma manera, no se observan discontinuidades en sus frecuencias espectrales lineales y ninguna discontinuidad en sus formantes en el dominio de la frecuencia, dentro del área de concatenación del diptongo /aU/. Las distorsiones de concatenación son mínimamente perceptibles al escuchar la palabra de forma auditiva.

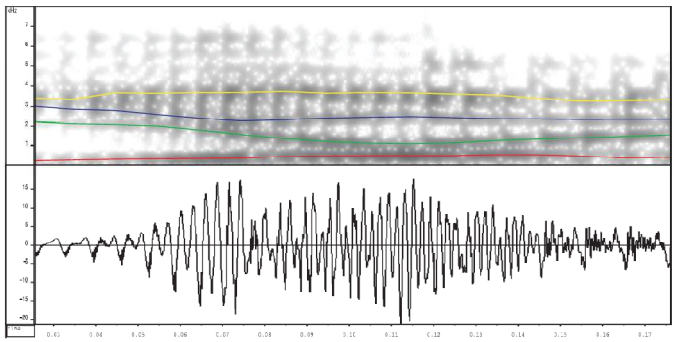

Fig. 7 OPC para el diptongo /OI/, palabra "join".

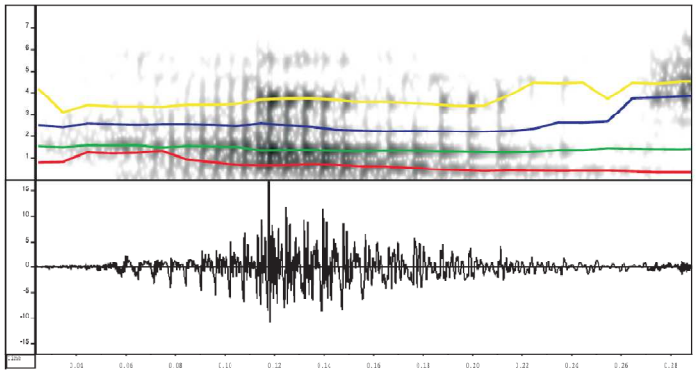

Fig. 8 OPC para el diptongo /aU/, palabra "house".

La optimización a través del método de OPC para la palabra “same” contenedora del diptongo /eI/ se ilustrada dentro de la siguiente Fig. 9. En mencionada figura se muestra que no existen discontinuidades algunas entre las frecuencias espectrales lineales ni discontinuidades en sus formantes

espectrales dentro del área de concatenación del diptongo /eI/ así como ninguna distorsión auditiva es percibida al escuchar la palabra sintetizada “same” contenedora del diptongo.

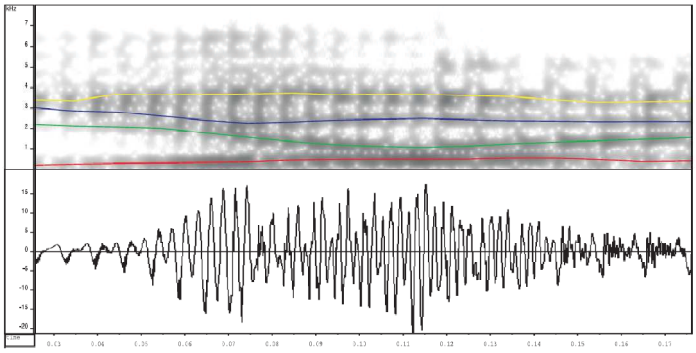

Fig. 9 OPC para el diptongo /eI/, palabra "same".

VI. Conclusiones

Se ha introducido un método de optimización del punto de concatenación (OPC) basado en PCA para la síntesis de voz. Esta optimización evita el post-procesamiento (suavizado) de la señal de voz sintetizada, ya que este método estima el punto óptimo de concatenación entre todas las unidades de voz dentro de un inventario. Logrando así evitar la aparición de artefactos y distorsiones en la voz sintetizada de un texto de entrada. El método OPC se deriva de la descomposición modal de los vectores de entrada de datos extraídos a través de las áreas de concatenación. OPC basa su funcionamiento en el análisis PCA donde estima las características principales de las unidades de voz al momento de estimar el punto de concatenación óptimo. Para ello, en las regiones frontera o áreas de concatenación, los periodos fundamentales de Tono de todas las posibles unidades candidatas son extraídas y asignadas (mapeadas) dentro del espacio característico del PCA. Posteriormente, cada combinación de concatenación es calculada en términos de la medida de distancia. Finalmente, los puntos de concatenación con la menor discontinuidad acumulada son seleccionados entre todas las posibles concatenaciones. Estos puntos representan los puntos de concatenación óptimos entre un conjunto de unidades de voz para el selector de unidades. El método propuesto se evaluó mediante el análisis de la contigüidad de la señal sintetizada mediante el análisis de espectrograma, frecuencias espectrales lineales, así como pruebas auditivas mostradas en las Figuras 7, 8 y 9. El análisis de cada uno de los diptongos sintetizados no mostro discontinuidad espectral alguna así como ninguna discontinuidad en sus formantes en el área de concatenación de la señal de voz sintetizada. Por lo tanto, se concluye que el OPC puede ser considerado como un método de entrenamiento óptimo para determinar el punto de concatenación adecuado entre un conjunto de unidades de voz dentro de un selector de unidades de voz en un sintetizador de voz.

Referencias

[1] Luna-Garcia, H., Mendoza-Gonzalez, R., Gamboa-Rosales, H., Celaya-Padilla, J., Galvan-Tejada, C., Lopez-Monteagudo, F., Collazos-, C., Mendoza-Gonzalez, A..(2018). MENTAL MODELS ASSOCIATED TO VOICE USER INTERFACES FOR INFOTAINMENT SYSTEMS. DYNA, 93(3). 245. DOI: http://dx.doi.org/10.6036/8766

[2] Galván-Tejada C, López-Monteagudo F, Alonso-González O, Galván-Tejada J, Celaya-Padilla J, Gamboa-Rosales H, Magallanes-Quintanar R, Zanella-Calzada L (2018) A Generalized Model for Indoor Location Estimation Using Environmental Sound from Human Activity Recognition. ISPRS International Journal of Geo-Information 7(3), 81. doi:10.3390/ijgi7030081.

[3] Celaya-Padilla, J., Galván-Tejada, C., López-Monteagudo, F., Alonso-González, O., Moreno-Báez, A., Martínez-Torteya, A., Galván-Tejada, J., Arceo-Olague, J., Luna-García, H. and Gamboa-Rosales, H. (2018). Speed Bump Detection Using Accelerometric Features: A Genetic Algorithm Approach, Sensors 18(2): 443. Retrieved from http://dx.doi.org/10.3390/s18020443

[4] Nandal, A., Dhaka, A., Gamboa-Rosales, H., Marina, N., Galvan-Tejada, J., Galvan-Tejada, C., Moreno-Baez, A., Celaya-Padilla, J. and Luna-Garcia, H. (2018). Sensitivity and Variability Analysis for Image Denoising Using Maximum Likelihood Estimation of Exponential Distribution. Circuits, Systems, and Signal Processing, 37(9), pp.3903-3926.

[5] Gamboa Rosales, H. (2010). Optimale Bausteinauswahl in der Korpusbasierten Sprachsynthese. Dresden: TUD press.

[6] Nematollahi, M., Vorakulpipat, C. and Gamboa, H. (2017). Digital Watermarking. 1st ed. Springer Singapore:.

[7] Gamboa Rosales, H., Jokisch, O. and Hoffmann, R. (2006). SPECTRAL DISTANCE COSTS FOR MULTILINGUAL UNIT SELECTION IN SPEECH SYNTHESIS. Proc. International Conference on Speech and Computer (SPECOM), pp.270-273.

[8] Mohammad Ali Nematollahi, ChaleeVorakulpipat, and Hamurabi Gamboa Rosales, “Semifragile Speech Watermarking Based on Least Significant Bit Replacement of Line Spectral Frequencies,” Mathematical Problems in Engineering, vol. 2017, Article ID 3597695, 9 pages, 2017. https://doi.org/10.1155/2017/3597695.

[9] Mohammad Ali Nematollahi, Chalee Vorakulpipat, and Hamurabi Gamboa Rosales, “Optimization of a Blind Speech Watermarking Technique against Amplitude Scaling,” Security and Communication Networks, vol. 2017, Article ID 5454768, 13 pages, 2017. https://doi.org/10.1155/2017/5454768.

[10] Mohammad Ali Nematollahi, Chalee Vorakulpipat, and Hamurabi Gamboa Rosales, “Semifragile Speech Watermarking Based on Least Significant Bit Replacement of

Line Spectral Frequencies," Mathematical Problems in Engineering, vol. 2017, Article ID 3597695, 9 pages, 2017. https://doi.org/10.1155/2017/3597695.

[11] Nematollahi, M.A., Al-Haddad, S.A.R., Doraisamy, S. et al. Natl. Acad. Sci. Lett. (2016) 39: 197. https://doi.org/10.1007/s40009-016-0430-8

[12] Carlos E. Galván-Tejada, Jorge I. Galván-Tejada, José M. Celaya-Padilla, et al., "An Analysis of Audio Features to Develop a Human Activity Recognition Model Using Genetic Algorithms, Random Forests, and Neural Networks," Mobile Information Systems, vol. 2016, Article ID 1784101, 10 pages, 2016. https://doi.org/10.1155/2016/1784101.

[13] Luna-Garcia, H., Mendoza-Gonzalez, R., Gamboa-Rosales, H., Celaya-Padilla, J., Galvan-Tejada, C., Lopez-Monteagudo, F., Collazos-, C., Mendoza-Gonzalez, A..(2018). MENTAL MODELS ASSOCIATED TO VOICE USER INTERFACES FOR INFOTAINMENT SYSTEMS. DYNA, 93(3). 245. DOI: http://dx.doi.org/10.6036/8766

[14] Nematollahi, M.A., Vorakulpipat, C., Gamboa-Rosales, H. et al. Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. (2017) 87: 433. https://doi.org/10.1007/s40010-017-0371-8

[15] Nematollahi, M.A., Gamboa-Rosales, H., Martinez-Ruiz, F.J. et al. Multimed Tools Appl (2017) 76: 7251. https://doi.org/10.1007/s11042-016-3350-1

[16] Luna-García, H., Mendoza-González, R., Gamboa-Rosales, H., Celaya-Padilla, J., Galván-Tejada, C., López-Monteagudo, F., Collazos-, C., Mendoza-González, A..(2018). MENTAL MODELS ASSOCIATED TO VOICE USER INTERFACES FOR INFOTAINMENT SYSTEMS. DYNA, 93(3). 245. DOI: http://dx.doi.org/10.6036/8766

[17] García-Hernández A, Galván-Tejada C, Galván-Tejada J, Celaya-Padilla J, Gamboa-Rosales H, Velasco-Elizondo P, Cárdenas-Vargas R (2017) A Similarity Analysis of Audio Signal to Develop a Human Activity Recognition Using Similarity Networks. Sensors 17(11), 2688. doi:10.3390/s17112688.

[18] Hossmar, Maria. Master Thesis: "Emotional Prosody for Speech Synthesis". Dresden University of Technology, 2009.

[19] Kotnik, B., "Determination of characteristic points inside the glottal cycle", Technical report, University of Maribor, Faculty of Electrical. Engineering and Computer Science, 2005.

[20] Galván-Tejada C, López-Monteagudo F, Alonso-González O, Galván-Tejada J, Celaya-Padilla J, Gamboa-Rosales H, Magallanes-Quintanar R, Zanella-Calzada L (2018) A Generalized Model for Indoor Location Estimation Using Environmental Sound from Human Activity Recognition. ISPRS International Journal of Geo-Information 7(3), 81. doi:10.3390/ijgi7030081.

[21] Upadhyay, N. & Rosales, H.G. Natl. Acad. Sci. Lett. (2018) 41: 15. https://doi.org/10.1007/s40009-017-0597-7