Concatenation point optimization by principal component analysis

Hamurabi Gamboa-Rosales, Roberto Olivera-Reyna, Salvador Ibarra-Delgado
Osvaldo Vite-Chávez, Reynel Olivera-Reyna, José Ismael De La Rosa-Vargas

Optimización de puntos de concatenación por análisis de componentes principales

Recibido: mayo 12, 2012
Aceptado: junio 2, 2012

Palabras clave: síntesis de voz; selección de la unidad; concatenación de punto; concatenación de segmento; percepción de discontinuidad.

Abstract:
Generally, concatenative speech synthesis systems provide a considerable synthesis quality since the criteria for unit selection methods have been optimized. However, the level of synthesis quality depends on the adequate position of the concatenation points of all acoustic units that have to be concatenated. The position of the concatenation points heavily determines the grade of mismatch and distortion human perception in a synthesized waveform. Therefore, we present a concatenation point optimization (CPO) algorithm based on Principal Component Analysis (PCA) that establishes an optimal concatenation point between any two matching acoustic units in a given inventory and reduces the distort human perception in Text-To-Speech Synthesis (TTS) Systems. The algorithm extracts data frames referring to a concatenation point and transforms them, using PCA, into a particularly framework, preserving the relevant properties of the waveform. Afterwards, we determined the optimal concatenation point by a task optimization. Experimental evaluations characterize the behavior of the proposed concatenation point optimization method and emphasizes its viability.

Keywords: speech synthesis; unit selection; concatenation point; segment concatenation; discontinuity perception

CONCATENATIVE speech synthesis has been used on TTS systems over the last years [1]. By this approach, the acoustic units of the inventory cover a big variety of phonetic and prosodic language features. Consequently, the unit selection extracts the best unit sequence from the entire inventory to synthesize an input text by minimizing the mismatches and distortions of the concatenated acoustic units. In most of the cases, the units searched by the unit selection are extracted from different phonetic contexts and present discontinuities in spectral shape as well as phase mismatches at concatenation boundaries. Additionally, the extracted units typically consist of variable-length phoneme, diphone or triphone sequences, which produce a larger number of concatenation points in a synthesized waveform. Because such discontinuities and mismatches deteriorate significantly the speech synthesis quality by the concatenation of acoustic units, the development of an optimal concatenation approach has become a hard task in speech synthesis[1]. Normally, distortion human perception of the join between acoustic units is

estimated in unit selection process by calculating the concatenation cost. This is calculated as the weighted sum of n concatenation sub-costs such as FFT-Spectrum, LPC-coefficients, linear spectral frequencies (LSF) coefficients, frequency cepstral coefficients (MFCC), or multiple centroid analysis (MCA) [4].

Concatenation algorithms still calculate a general concatenation point instead of finding an optimal concatenation point

However, they all are derived from Fourier signal analysis, and each distortion is related more or less significantly to the discontinuity measure at the spectral area between the fixed concatenation points of acoustic units [2]. The appropriated set up of the concatenation points produces a higher speech synthesis quality avoiding the appearance of artifacts between the concatenated units in a synthesized waveform. However, concatenation algorithms still calculate a general concatenation point instead of finding an optimal concatenation point for a given set of acoustic units. Therefore, we propose a methodical algorithm to obtain the optimal concatenation point between a set of candidate acoustic units via PCA and so to make the likelihood of a bad concatenation effectively small. PCA gives an alternative feature data extraction and provides a new analysis construction to characterize the acoustic mismatch between two units. Because PCA transform the feature data in a framework that contains the relevant properties in the concatenation area, the resulting CPO is more explicit when it comes to compare a set of concatenation candidate units against each other. We refer to this (off-line) concatenation point optimization for unit selection. In the following section this paper provides a general description of the PCA framework and feature data extraction. Then, we analyze in more detail the PCA modal decomposition and its principal characteristics for the CPO, and we give an overview on the CPO method. Finally, the experimental analyses are reported concerning the concatenation point estimation in a TTS-System.

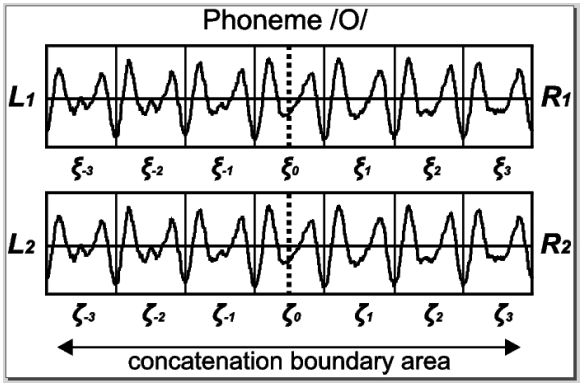

Figura 1. Centered Pitch Period Notation $(K = 3)$.

PCA FRAMEWORK

Firstly, we consider a diphone style concatenative-based TTS-System. It means that all matching diphones starting or ending within the phoneme /O/ are collected among the set of recorded utterances in a given inventory. So, we can concentrate on finding the acoustic units and their optimal concatenation points within /O/ that reduce the mismatch and distortion when they are concatenated. In Fig. 1 two such acoustic units are shown, denoted by $L_1 - R_1$ and $L_2 - R_2$, where L_1 represents the first-half and R_1 shows the contiguous second half of the left-hand side diphone, and L_2 illustrates the first-half and R_2 shows the contiguous second half of the right-hand diphone [1]. So, we focused on finding the optimal concatenation point between L_1 and R_2, whose unit is not available in the inventory.

Let $\xi_{-K+1}, ..., \xi_0, ..., \xi_{K-1}$ and $\zeta_{-K+1}, ..., \zeta_0, ..., \zeta_{K-1}$ denote the $2K-1$ centered pitch periods associated with the concatenation area of $L_1 - R_1$ and $L_2 - R_2$ respectively as it is proposed by [3]. Additionally, the interior boundaries between $L_1 - R_1$ and $L_2 - R_2$ fall exactly in the middle of ξ_0 and ζ_0 as it is shown in Fig. 1. The pitch marking for the units is obtained for the voice and voiceless speech units via [5]. Consider that there are M_1 units, like $L_1 - R_1$, and M_2 units, like $L_2 - R_2$, with a concatenation point within /O/ in the entire unit inventory. Additionally, the centered pitch periods are estimated for everyone of these units with the methods mentioned above. This implies the search for an optimal concatenation point across the $(2K-1)(M_1 \times M_2)$ centered pitch periods, assuming $M_1 \times M_2 = M$ concatenation combinations and N as the maximum possible number of samples between the pitch periods per unit. Also, sym-

metrical zero-pad is applied if it is necessary, as well as appropriate windowing for all centered pitch periods N. The result is a $((2K-1)M \times N)$ matrix X with the elements $x_{i,j}$, where x_i represents the centered pitch periods and the x_j column represents the slice of time samples [6]. Further, we transform the data input matrix X by performing PCA [7]. PCA decomposes a data set of mixed signals into a data set of uncorrelated signals. In terms of moments, this implies that PCA finds a matrix that transforms the input data $X(x_1^1, x_2^2, ..., x_{(2K-1)M}^N)$ with a probability density function (pdf) $p(x_1^1, x_2^2, ..., x_{(2K-1)M}^N)$ into a set of uncorrelated signals $Y(y_1^1, y_2^2, ..., y_{(2K-1)M}^N)$ as it is showed in equation (1).

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

We propose a methodical algorithm to obtain the optimal concatenation point between a set of candidate acoustic units via PCA

Let μ_x be the mean of the population input matrix X and matrix A consist of the eigenvectors of the covariance matrix of the input matrix X as it is shown in the rows of the matrix A_T in Fig. 2. Mean subtraction is an integral part of the solution towards finding a principal component basis, which minimizes the mean square error of data approximation. Afterwards, it is possible to reconstruct the input matrix X by using Y and the orthogonal property $A^T = A^{-1}$:

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

where A is the $(2K-1)M \times (2K-1)M$ orthogonal basis for the PCA space L spanned by the $(2K-1)M$-dimensional a_i's. Y can be seen as the coordinates in the orthogonal base. The input matrix X is projected on the coordinate axes defined by the orthogonal basis. In this form the input matrix X (which contains all candidate units) is represented by a linear combination of the orthogonal basis vectors for the space L with the dimension $(2K-1)M$. This can be described as a rank-M decomposition, which defines a mapping between the set of centered pitch periods and the set of $(2K-1)M$-dimensional feature vectors $\tilde{a}_i = a_i(Y + \mu_x)$.

The proposed feature extraction methodology gives a global overview of the component vectors that are interacting directly at the concatenation area for the phoneme /O/ as it is illustrated in Fig. 2. In fact, the CPO of the candidate feature vector units is determined by the overall features observed in the pitch periods within the PCA framework, in contrary to an analysis restricted to a specific instance or frequency domain. Therefore, the join of two vectors a_ξ and a_ζ in a new feature space (PCA transformation) can reflect a high degree of similarity, and thus potentially a degree of mismatch and distortion could be also estimated.

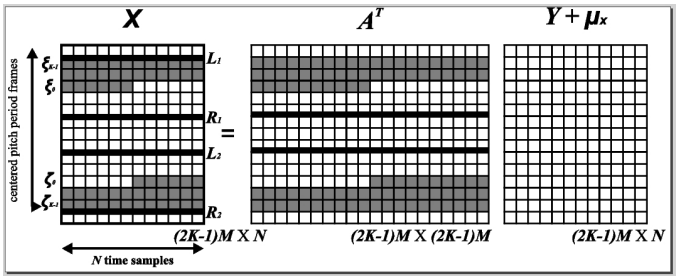

Figura 2. PCA Representation of the Input Matrix X.

CONCATENATION POINT ANALYSIS

Concatenation point optimization must be carried out over all M possible concatenation boundaries between the $L_1 - R_1$ and $L_2 - R_2$ unit candidates. CPO must search the concatenation points across the concatenation boundary areas of all combinations of $L_1 - R_2$ units so that the M concatenations present minimal distortion when they are concatenated. Afterwards, the units that present the minimum distortion at the optimized concatenation point are selected. To achieve this task, the possible concatenation points from the units have to be represented in the space L.The PCA transformation comprises the vector a_{ξ_k} and a_{ζ_k}, which represent the centered pitch periods of ξ_k and ζ_k for $-K+1 \leq k \leq K-1$. Consider now the candidate concatenation between $L_1 - R_2$ as it is shown in the shaded area in Fig. 2. This concatenation point can be described as $\xi_{-K+1}, ..., \xi_1, \phi_0, \zeta_1, ..., \zeta_{K-1}$, where ϕ_0 illustrates the concatenated centered period. It is composed by the left half of ξ_0 and the right half of ζ_0 and can be represented in the space L of the PCA transformation of the feature input data matrix X as followed:

$$\tilde{a}_{\xi_{-K+1}}, ..., \tilde{a}_{\xi_1}, \tilde{a}_{\phi_0}, \tilde{a}_{\zeta_1}, ..., \tilde{a}_{\zeta_{K-1}} \quad (3)$$

where $\tilde{a}_{\xi}$ represent the left half side of the unit L_1 in the matrix A and $\tilde{a}_{\zeta_1}$ represents the right side of the unit R_2 in the matrix A. However, $\tilde{a}_{\phi_0}$ does not have a representation in the input matrix X. It can be calculated by computing ϕ_0 as an additional row of the input feature matrix X as it is shown in the following formula (4):

$$\phi_0 = a_{\phi_0} R = \tilde{a}_{\phi_0} \quad (4)$$

where the $(2K-1)M$-dimensional vector a_{ϕ_0} is introduced as an additional row in the matrix A, and R is the operation corresponding to $Y + \mu_x$. So, the concatenation vector $\tilde{a}_{\phi_0} = a_{\phi_0} R$ represents the ϕ_0 in the space L. Once the $\tilde{a}_{\phi_0}$ is estimated, the mismatch and distortion of the units to be concatenated is measured as a cumulative closeness difference between the vectors composing the two units segments $\tilde{a}_{\xi_1}$ and $\tilde{a}_{\zeta_1}$. It is achieved with the help of the closeness measure between two individual vectors as it is proposed by [3],[6]:

$$s(\tilde{a}_k, \tilde{a}_l) = cos(a_k R, a_l R) = \frac{a_k R R^T a_l^T}{\|a_k R\| \, \|a_l R\|} \quad (5)$$

where $1 \leq k, l \leq (2K-1)M$. Taking the shorthand notation from [6] results in:

$$\tilde{s}(a_{\alpha_{-k}}, a_{\alpha_0}, a_{\alpha_k}) = \frac{s(\tilde{a}_{\alpha_{-k}}, \tilde{a}_{\alpha_0}) + s(\tilde{a}_{\alpha_0}, \tilde{a}_{\alpha_k})}{2} \quad (6)$$

Finally, the average similarity between the boundary vectors of two concatenated units can be described as followed [6]:

$$d(V_1, V_2) = \sum_{k=1}^{K-1} 2\tilde{s}(a_{\xi_k}, a_{\phi_0}, a_{\zeta_k}) - \quad (7)$$

$$\tilde{s}(a_{\xi_k}, a_{\xi_0}, a_{\xi_{-k}}) - \tilde{s}(a_{\zeta_{-k}}, a_{\zeta_0}, a_{\zeta_k})$$

This dissimilarity estimation corresponds to the measure of the trajectory difference before and after concatenation across the entire concatenation boundary area. It can be tested, when the difference measure between two contiguous units in the database is calculated, i.e. the ξ's are equal to ζ's. It means that $\phi_0 = \xi_0 = \zeta_0$ and $d(V_1, V_2) = 0$ if, and only if, $L_1 = R_2$. The closer the difference measure is to zero, the more adequate the concatenation point is between two specific units.

CONCATENATION POINT OPTIMIZATION

Once the PCA framework for the CPO is specified, the following step is to develop a procedure to find out the concatenation point and the corresponding units that present the least distortion when they are concatenated. The idea is to compute the accumulative $d(V_1, V2)$ distance focusing on every possible concatenation point in the area $[-K, K]$, using the space L,which is associated to the concatenation area as it is shown in the flowchart of Fig 3.

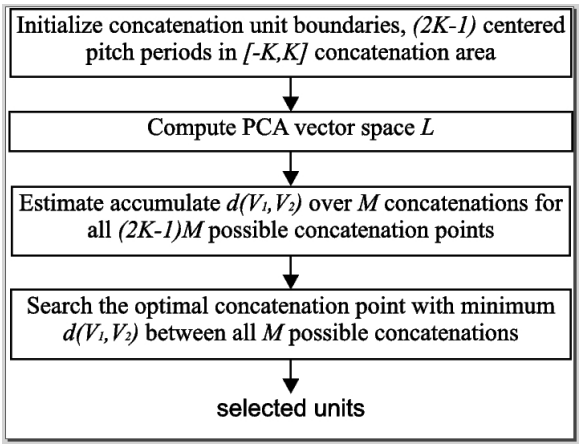

Figura 3. Concatenation Point Optimization.

Firstly, the initialization step establishes the centered pitch period data frames $2K-1$ in the concatenation area of $[-K, K]$. Then, we derive the centered pitch period data frames into the space L by utilizing PCA. The outcome provides $(2K-1)M$ feature vectors in the space L, with the same number of potential concatenation points. Afterwards, we compute the accumulative discontinuity associated to every concatenation point for all M candidate concatenations. Finally, the concatenation point and the corresponding units that present the minimum accumulative discontinuity are selected for unit selection process. Therefore, we resume the PCA framework and CPO procedures. Both procedures are depicted in detail in Fig. 4. The centered pitch period data frames of the units are extracted from the concatenation boundary areas of the units. Then they

are derived into a L space by using PCA. Afterwards, the concatenation point with the least accumulative discontinuity $d(V_1, V_2)$ is selected between all possible concatenations. This point represents the optimal concatenation point (offline) for unit selection.

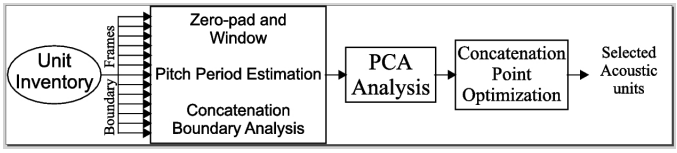

Figura 4. PCA Framework and CPO Procedures.

EXPERIMENTAL RESULTS

Syrdal mentioned that a reliably higher discontinuity-detection rate is observed in diphthongs than in monophthong vowels [10], [9]. Therefore, we focused on the analysis of the concatenation mismatch of the proposed CPO method by synthesizing the English diphthongs /eI/, /OU/, /aI/, /aU/, and /OI/. Dress TTS system and the "TC-STAR" English speech database were used to synthesize the mentioned diphthongs [8]. The synthesis of the diphthongs was achieved by joining two diphones resulting in a specific diphthong inside a word, when they are concatenated. Firstly, we extracted all the units with the left or right boundary falling in the middle of the corresponding phoneme of the diphthongs from the entire inventory. There were a total of 2902 instances for /eI/, 2094 instances for /OU/, 3373 instances for /aI/, 1496 instances for /aU/, and 644 units for /OI/. Additionally, we extracted $K = 5$ pitch periods on the left and right side of the the concatenation boundary, giving $2K - 1 = 9$ centered pitch periods. Afterwards, we followed the procedure described in the previous sections for the concatenation point optimization by synthesizing the diphthong /eI/ inside of the word "same". Therewith, we obtained an input data matrix X of a size of (645 $\times$ 128), which is composed of 28 and 47 units for the left and right sides of all possible concatenation combinations, respectively. Then, we applied the concatenation point optimization for the diphthong /OU/ inside the word "hope". So, an input data matrix X (1199 $\times$ 128) with 36 and 96 units for the left and right sides was obtained. The concatenation point optimization for the diphthong /aI/ inside the word "Kite" delivers an input data matrix X (3553 $\times$ 128X) with 45 and 372 units for the left and right sides. In the same form, the concatenation point optimization for the diphthong /aU/ inside the word "house" delivers an input data matrix X (1296 $\times$ 128), which is integrated for 83 and 61 units for the left and right sides of the possible concatenation combinations respectively. Finally, we obtained a input data matrix X for the diphthong /OI/ inside the word "join" of a size of(144 $\times$ 128), which is integrated for 5 and 11 units for the left and right side. In Fig. 5 the concatenation point optimization for the word "join" containing the diphthong /OI/ is shown. We can observe the speech signal,

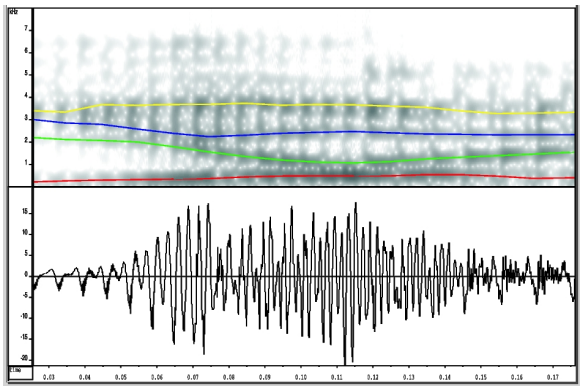

Figura 5. COP for the Diphthong /OI/, word: join.

spectrogram, and the formants frequencies of the synthesized word "join" in Fig. 5. We notice any spectral discontinuities and any relevant formant discontinuities in the concatenation area of the diphthong /OI/ in the synthesized speech signal. Additionally, any concatenation distortion is perceived by listening this word. Concatenation point optimization for the word "house" containing the diphthong /aU/ is shown in Fig. 6. In the same way, we can see any significantly spectral discontinuities and any important formant discontinuities in the concatenation area of the diphthong /aU/ in Figure 6. Concatenation distortions are almost unperceived by listening this word. Concatenation point optimization for the word "same" containing the diphthong /eI/ is shown in Fig. 7. Any significant spectral discontinuities and any formant discontinuities in the concatenation area of the diphthong /eI/ appear in Fig. 7. Additionally, any concatenation distortion is perceived by listening this word.

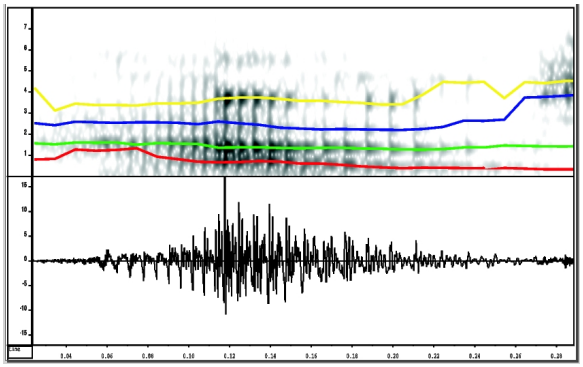

Figura 6. CPO for the Diphthong /aU/, word: house.

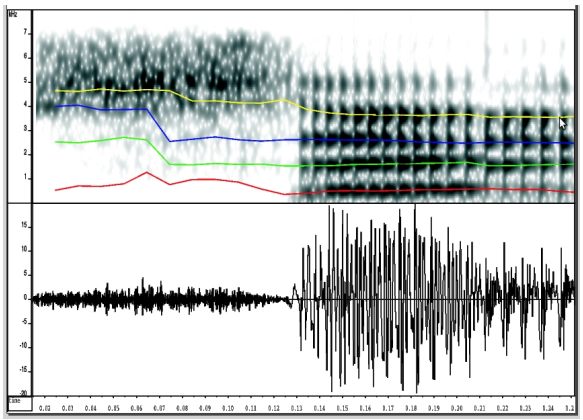

Figura 7. CPO for the Diphthong /eI/, word: same.

CONCLUSIONS

We have introduced the concatenation point optimization for unit selection process in concatenative-based speech synthesis via PCA. This optimization avoids the post-processing (smoothing) of the synthesized speech signal, because this method estimates the best concatenation point between speech units avoiding possible mismatches and distortions. The CPO is derived from the modal decomposition (PCA) of data vectors collected across the entire boundary concatenation area. CPO bases its function on PCA feature analysis. It preserves those properties of the acoustic units that are important to consider in the concatenation point optimization. Around a given concatenation boundary region, the pitch periods are extracted from all possible candidate units and they are mapped onto a PCA feature space. Then, each concatenation combination is calculated in terms of the distance measure exposed by [3]. Finally, the concatenation point with the least accumulative discontinuity is selected between all possible concatenations. This point represents the optimal concatenation point between a set of speech units for unit selection. The proposed method was evaluated by analyzing the contiguity of the synthesized speech signal by using the spectrogram as it is shown in Fig. 5, 6, and 7. The analysis of everyone of the synthesized diphthongs depicted any relevant spectral discontinuities and any important formant discontinuities in the concatenation area of the synthesized speech signal. Therefore, we conclude that the CPO can be considered as an optimal training method to determine the concatenation point between a set of speech units for the unit selection process.

Bibliografía

[1] Hunt, A.J. and Black, A.W., "Unit selection in a concatenative speech synthesis using a large speech database", in Proc. ICASSP, pp. 373-376, 1996.

[2] Vepa, J., King, S. and Taylor, P., "Objective distance measures for spectral discontinuities in concatenative speech synthesis", In ICSLP, Denver, USA, 2002.

[3] Bellegarda, J. R., "A Novel Dicontinuity Metric for Unit Selection Text-to-speech Synthesis", in Proc. 5th ISAC Speech Synth. Workshop, Pittsburg, PA, pp. 13338, June 2004.

[4] Corwe, A. and Jack, MA., "Globally optimizing formant tracker using generalized centroids", Electronic Letters, Vol 23, No. 19, pp 1019-1020 Beijing, China, 1987.

[5] Hussein, H. and Jokisch, O., "Hybrid electroglottograph and speech signal based algorithm for pitch marking", In INTERSPEECH-2007, pp. 1653-1656.

[6] Bellegarda, J.R., "LSM-Based Boundary Training for Concatenative Speech Synthesis",in Proc. ICASSP, Toulouse, France, May 2006.

[7] Jolliffe, I.T., "Principal Component Analysis", Springer-Verlag New-York, 1986. Survey, PCA.

[8] Gamboa Rosales, H. and Jokisch, O., "Korpus-Dress1 - Korpusbasierte Konkatenative Sprachsynthesesysteme", In Proc 18. Konferenz Elektronische Sprachsignalverarbeitung, Cottbus, Germany, pp. 115-122, 2007.

[9] Gamboa Rosales, H., Jokisch, O. and Hoffmann, R., "Spectral distance costs for multilingual unit selection in speech synthesis", In Proc. SPECOM'2006, St. Petersburg, Russia, pp. 270-273,2006.

[10] Syrdal, A.K., "Phonetic effects on listener detection of vowel concatenation", in Ptoc. Eurospeech, Aalborg, Denmark, 2001.

Acerca del autor o autores

Hamurabi Gamboa Rosales es ingeniero en Comunicaciones y Electrónica egresado de la Universidad Autónoma de Guadalajara en el año 2000. Obtuvo el grado de Maestro en Ciencias con especialidad en Ingeniería Eléctrica en el área de Procesamiento Digital de Señales (PDS) en 2003 por parte de la Universidad de Guanajuato. Posteriormente, obtiene el grado de Doctor en Ciencias con especialidad en Ingeniería Eléctrica en el área de Procesamiento de Señales en 2010, en el Institute of Acoustics and Speech Communication de la Universidad Tecnológica de Dresde, Alemania. Se incorporó a la Universidad Autónoma de Zacatecas (UAZ) en agosto de 2003, como profesor-investigador de tiempo completo adscrito a la Unidad Académica de Ingeniería Eléctrica. Actualmente es miembro del SNI como Candidato de 2012 a 2014. Trabajó recientemente en la industria Europea con permiso de la UAZ en voiceINTERconnect GmbH Dresden, Sajonia, Alemania y en Nuance Communications International, Merelbeke, Flandes, Bélgica.