Computación y Sistemas Vol. 3 No. 2 pp. 98 - 104
© 1999, CIC - IPN. *ISSN 1405-5546* Impreso en México

Entrenamiento de un Reconocedor Fonético de Dígitos para el Español de México usando el CSLU Toolkit

N. Munive, A. Vargas, B. Serridge, O. Cervantes e I. Kirschning
Universidad de las Américas-Puebla
Grupo de Investigación en Tecnología de voz Tlatoa[1]
Cholula, Puebla C.P.72820, México
Tel: (22) 29 26 23 fax: (22) 29 21 38
e-mail: {nora,alcira,ben,ocervan,ingrid}@mail.pue.udlap.mx

Artículo recibido el 10 de junio, 1999; aceptado el 27 de agosto, 1999

Resumen

En este trabajo se presenta el diseño de un reconocedor fonético de dígitos para el español hablado en México, implementado usando el Toolkit desarrollado por el Center for Spoken Language Understanding (CSLU) de Portland, Oregon.

Se discuten las principales herramientas y metodología del CSLU Toolkit con base en las cuales se diseñó el reconocedor. Se expone la creación de una base de datos de dígitos, fonéticamente balanceada, la cual fue empleada como datos de entrenamiento en el desarrollo del sistema. Finalmente, se presentan los resultados obtenidos en las distintas pruebas efectuadas para evaluar el reconocedor.[2]

Palabras Clave

Reconocimiento de voz; español mexicano; diseño de corpus de texto y voz; CSLU *Toolkit*.

Introducción

En este artículo se expone el desarrollo de un reconocedor fonético de dígitos de habla continua para el español hablado en México usando el CSLU *Toolkit*. Este amplio conjunto de herramientas y tecnologías apoyan la investigación, aprendizaje y desarrollo de sistemas interactivos que usan la voz como interfaz. El ambiente de programación que ofrece está estructurado modularmente y es muy flexible lo que hace posible integrar nuevos componentes. (Fanty, 1996)

Entre los trabajos realizados anteriormente en el área destacan los realizados por laboratorios especializados (MIT, CMU, OGI) principalmente para el idioma Inglés, Alemán y Japonés. (Hosom *et al.*, 98) (Shultz *et al.*, 97) En particular, el MIT ha empezado a diversificar sus sistemas a otros lenguajes, sin embargo la portabilidad de un lenguaje no es tan directa; se requiere de la participación de hablantes nativos que puedan entender y modelar los fenómenos lingüísticos y de esta manera lograr desarrollar sistemas robustos de lenguaje hablado. Por otro lado, España ha empezado a desarrollar sistemas (Tapias *et al.*, 1994) pero evidentemente el Español de España y México tienen diferencias significativas dialectales en la pronunciación y entonación que afectan el desempeño de los reconocedores.

La elaboración de este trabajo se llevó a cabo con base en la metodología propuesta por el CSLU para la creación de sistemas de reconocimiento de voz. El proceso de desarrollo del reconocedor se realizó en dos etapas; en la primera se creó un corpus de dígitos para el español el cual fue utilizado en la segunda etapa para la construcción de un clasificador fonético dependiente del contexto. En la primera sección de este trabajo se describe la creación de un corpus de dígitos.

[1]Miembro del CENTIA-UDLAP
[2]Este proyecto fue realizado gracias al apoyo prestado por NSF-Conacyt, con el No. CO-66-A9605.

Enseguida se expone la etapa de construcción de un clasificador basado en el enfoque de redes neuronales. Finalmente, se discuten los resultados obtenidos en la evaluación del reconocedor de dígitos.

1 Obtención del Corpus de Dígitos

El desarrollo de un corpus es una actividad clave en la creación de sistemas de lenguaje hablado. Para entrenar los modelos y crear reconocedores fonéticos de habla continua e independientes del locutor se requiere de la colección y transcripción de grandes cantidades de datos. Entre las actividades que involucran crear un corpus se encuentra el diseño del texto a grabarse, la creación de convenciones y documentación para etiquetar los datos y el desarrollo de herramientas que automaticen el etiquetado de los archivos.

1.1 Definición del Corpus de Texto

El diseño previo de las frases que conformarán un corpus permite un mejor modelado del vocabulario que se pretende reconocer. Algunas de las ventajas de un buen diseño (Cole *et al.*, 1996) son:

- asegura que los datos empleados en el entrenamiento sean similares a los que se pretenden reconocer y
- garantiza que se tengan suficientes muestras de los fonemas en cada contexto para entrenar la red.

Una vez definido el vocabulario, en este caso de dígitos, el siguiente paso es determinar las frases que conformarán el corpus de texto. Estas frases fueron diseñadas de manera que el contexto fonético entre cada par de dígitos esté balanceado. Cuando se trata del reconocimiento de palabras aisladas, es suficiente tener muchas muestras de cada palabra para asegurar que se tendrán suficientes ejemplos de los fonemas que forman esa palabra. La tabla 1 muestra los fonemas en la palabra "uno".

contexto izquierdo	fonema	contexto derecho
pausa	/u/	n
u	/n/	o
n	/o/	pausa

Tabla 1: Contextos fonéticos de la palabra "uno

Pero en el caso del reconocimiento de habla continua, también es importante tener muestras de contextos *entre* palabras. Para llevar a cabo esto, primero se agruparon las palabras del vocabulario de dígitos de acuerdo a su fonema inicial y terminal. Se obtuvieron los 2 grupos de contextos que se muestran en la tabla 2.

Contexto TERMINAL	
/o/	cero, uno, cuatro, cinco, ocho
/s/	dos, tres, diez, seis
/e/	siete, nueve

Contexto INICIAL	
/s/	cero, cinco, seis, siete
/u/	uno
/d/	dos, diez
/t/	tres
/k/	cuatro
/n/	nueve
/o/	ocho

Tabla 2: Contextos fonéticos entre palabras.

Posteriormente se generaron todas las combinaciones posibles entre los contextos obteniendo 28 grupos de pares de palabras que comparten la misma combinación de fonema terminal con fonema inicial.

Por ejemplo, los pares de palabras que resultan en la combinación /e/ → /d/ son:

```
siete-dos, siete-diez

nueve-dos  nueve-diez
```

Tomando la misma cantidad de ejemplos de cada grupo, se obtuvieron las 40 frases que fueron empleadas en la colección de datos para crear el corpus de voz (Munive and Vargas, 1997). Cabe mencionar que cada frase está compuesta por una cadena de 6 dígitos.

1.2 Colección de Datos

El corpus de dígitos consiste en una colección de pronunciaciones grabadas de 50 personas (25 hombres y 25 mujeres). La mayoría de las personas son estudiantes universitarios procedentes de Puebla y del Distrito Federal pero también hay representantes de varias regiones del país. La tabla 3 muestra la población de locutores según su procedencia.

Las sesiones de grabación se llevaron a cabo en una oficina de 4.20 x 2.40 m. de dimensión sin aislamiento completo del ruido. El equipo utilizado consiste en una computadora PC Gateway 2000 con procesador Pentium, memoria de 2 Gb. en disco duro, tarjeta de sonido *Sound Blaster*, sistema operativo Solaris versión 2.51 y un micrófono diadema con filtro de ruido.

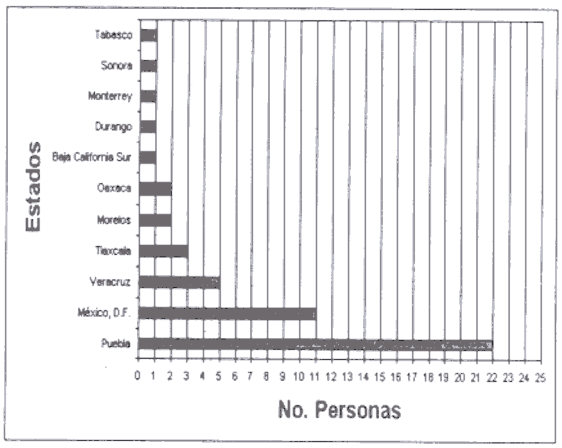


Tabla 3: Población de locutores por procedencia.

La señal de voz fue capturada a una frecuencia de muestreo de 8000 Hz. Se usó el CSLU *Toolkit* para controlar el diálogo con los locutores y cada persona leyó las mismas 40 frases del corpus de texto.

1.3 Transcripción Fonética

La transcripción o etiquetado de los archivos de voz se realizó primero a nivel de palabras no alineadas con respecto al tiempo y después a nivel de fonemas alineados con respecto al tiempo (Lander, 1996). La primera transcripción consistió en el etiquetado textual de lo que se pronunció, que en la mayoría de los casos corresponde con cada texto leído. Por ejemplo:

```
seis diez dos ocho siete nueve
```

La segunda transcripción se refiere al etiquetado fonético que consistió en identificar los fonemas pronunciados por el locutor con sus respectivos tiempos de inicio y terminación. El nombre del fonema se especificó usando los símbolos del alfabeto *Worldbet* (Hieronymus, 1993). Un ejemplo, de la palabra "dos", se muestra en la tabla 4, donde los valores de inicio y fin representan la duración del fonema en milisegundos.

fonema	inicio (mseg)	fin (mseg)
dc	462	506
d	506	527
o	527	701
s	701	817

Tabla 4: Ejemplo de etiquetas fonéticas.

Las etiquetas fonéticas fueron colocadas de forma manual para los primeros 10 locutores usando la herramienta gráfica *Lyre*, que forma parte del CSLU *Toolkit*. Usando este corpus etiquetado, se entrenó una primera versión del reconocedor, siguiendo el proceso que se describirá en la siguiente sección.

Después, empleando este reconocedor, se etiquetó automáticamente el resto del corpus al aplicar la técnica conocida con el nombre de *forced-alignment* (Hosom *et al.*, 1996). Esta técnica utiliza un reconocedor previamente entrenado para determinar el tiempo de inicio y terminación de los fonemas que se pronunciaron en un archivo de voz.

Normalmente, el reconocedor genera un conjunto de palabras reconocidas, pero además puede generar la lista de fonemas que se pronunciaron, forzándolo en cada pronunciación a obedecer una gramática que solo permite la secuencia de palabras que se tienen en la transcripción textual.

Dado que se cuenta con las transcripciones textuales de las pronunciaciones, se pudo usar el reconocedor para etiquetar el resto de las frases del corpus de forma automática, lo que ahorra bastante tiempo, ya que hacerlo manualmente es una tarea tediosa que consume mucho tiempo.

En esta etapa, finalmente se obtuvo un corpus de dígitos de 1,933 archivos de voz, cada uno con su transcripción textual y fonética. Esto es, 1 hr y 45 min de habla continua.

El corpus de dígitos fue dividido de manera aleatoria en los siguientes grupos:

60% para entrenamiento,
20% para desarrollo y
20% para evaluación final del reconocedor.

Los datos de entrenamiento se usan para entrenar el clasificador fonético basado en redes neuronales. En este conjunto de datos es importante tener suficientes muestras de cada unidad fonética para asegurar que la red neuronal aprenderá las características generales de los datos. El grupo de datos de desarrollo se emplea para evaluar el desempeño de la red en cada iteración de entrenamiento usando datos que no estén en el conjunto de entrenamiento. Por último, con el conjunto de datos de prueba se evalúa el nivel de reconocimiento final de la red neuronal que haya obtenido el mayor desempeño en la etapa de desarrollo.

2 Desarrollo del Clasificador

Una vez desarrollada la base de datos, se procedió a la construcción de un clasificador fonético dependiente del contexto para el vocabulario de dígitos. Esto se llevó a cabo siguiendo la metodología propuesta por el CSLU (Hosom and Cole, 1997), la cual consiste en:

- diseño del clasificador,
- entrenamiento de la red neuronal,
- elección de la mejor red con base en los datos de desarrollo y
- evaluación de la mejor red usando los datos de prueba.

2.1 Diseño del Clasificador

El sistema está basado en el reconocimiento de fonemas pero éstos varían mucho dependiendo de su contexto (i.e. sus fonemas vecinos). En el habla los fonemas están siempre influidos por los fonemas anterior y posterior, ya que los órganos articulatorios que producen los sonidos se encuentran en constante movimiento y no pueden cambiar instantáneamente de una posición a otra. Se ha comprobado que el fonema de la izquierda tiene más efecto al lado izquierdo del próximo fonema que al lado derecho y viceversa (Goldenthal, 1994).

Para tomar en cuenta estos efectos coarticulatorios, se puede dividir cada fonema en 3 partes, en donde la parte izquierda depende del contexto izquierdo, la parte derecha del contexto derecho y la parte central es independiente del contexto (Hosom and Cole, 1997). Así, con un conjunto de **N** fonemas, existen $2N^2+N$ posibles unidades distintas. Sin embargo, no todos los fonemas se dividen en tres partes, sólo si se considera necesario, dependiendo ello de la influencia que ejercen los fonemas vecinos.

Con el objeto de simplificar y reducir el número de unidades, los fonemas pueden además ser agrupados según sus características, ya sea de acuerdo con la manera de articulación, el lugar de articulación o si es conveniente una combinación de ambas. A cada grupo de fonemas se le denomina una clase o contexto general. Por ejemplo, se puede considerar a todos los fonemas nasales (/n/, /m/, /nj/) como un solo tipo de contexto o clase.

Cuando se habla del diseño del clasificador, se refiere a la división de fonemas en partes y a la agrupación de contextos en clases. Las tablas 5 y 6 muestran el diseño del reconocedor de dígitos.

No. partes	FONEMAS
1	tS tSc dc k kc n N s t tc
2	d r V
3	a e i o u w

Tabla 5: División de fonemas en partes.

CLASE	FONEMAS
$pau	tc kc tSc pau
$fnt	e i
$bck	o u w
$mid	a
$nas	n N
$fri	s tS
$ret	r
$lab	V
$vel	k
$wav	dc d t

Tabla 6: Clases generales de contexto.

1.2 Entrenamiento y Elección de la Mejor Red

Una vez que se define lo que se desea clasificar, en este caso categorías fonéticas, se procede al entrenamiento del clasificador. Éste es una red neuronal de arquitectura *feed forward* de tres niveles que emplea el algoritmo de aprendizaje *back-propagation* estándar con 130 nodos de entrada, 200 nodos ocultos y un nodo de salida por cada categoría fonética que se desea reconocer.

El proceso de entrenamiento de la red neuronal consistió en varios pasos. Primero, se divide la señal en frames y se calculan los vectores de características para cada uno. Estos vectores se obtienen a partir del espectro de la señal de voz y reflejan el conjunto de características relevantes de los fonemas. Para obtenerlos, se reduce un número determinado de muestras de la señal de voz a un conjunto de coeficientes que representan las concentraciones de energía y anchos de frecuencia de la señal. Para esto, se utilizó la técnica de procesamiento de señales MFCC (Mel frecuency Cepstral Coefficients) (Schalkwyk *et al.*, 1996).

Después, se introducen los vectores a la red neuronal que se encarga de clasificarlos en categorías fonéticas, ajustando los pesos gradualmente hasta encontrar aquellos similares a los que se desea obtener. Este proceso es iterativo, es decir se entrena un cierto número de veces suficiente para alcanzar un mínimo de error. Esta etapa tiene como objetivo que el clasificador aprenda las características esenciales de aquello que se desea reconocer.

La siguiente etapa es la de desarrollo que consiste en determinar en cuál iteración se obtuvo la red con el desempeño más alto. Para esto, se evalúa el nivel de error de reconocimiento a nivel de palabra alcanzado en cada iteración. Las palabras reconocidas se obtienen con el algoritmo de bús-

queda Viterbi (Hosom *et al.*, 1997) el cual utiliza un modelo de pronunciación para determinar la secuencia de palabras más probable. Para calcular el nivel de reconocimiento en

$$E = \frac{S + I + D}{N} * 100$$

términos del grado de error generado se aplica la siguiente fórmula:

en donde N es el número total de palabras en el conjunto de prueba, S es el número de substituciones, I el número de inserciones y D el número de supresiones. Esta fase identifica la red que está mejor preparada para reconocer las características generales y es la que se utilizará en la última etapa de prueba.

En total, se desarrollaron tres diferentes clasificadores, entrenados con el conjunto de datos para entrenamiento del corpus de dígitos etiquetado aplicando la técnica de *forced-alignment:*

- clasificador independiente del contexto,
- clasificador dependiente del contexto y
- clasificador dependiente del contexto agrupando los fonemas en clases generales.

Para el caso más simple, en el sistema independiente del contexto los fonemas no se dividieron en partes. En los otros dos casos, dependientes del contexto, cada fonema se dividió en una, dos o tres partes para tomar en cuenta el efecto de coarticulación que tienen sobre ellos los fonemas vecinos. En el tercer caso, además de la división de fonemas, se agruparon los contextos en clases generales.

Los tres sistemas fueron entrenados con el mismo conjunto de datos formado por 1158 frases (secuencias de 6 dígitos). El conjunto de desarrollo consistió de 375 frases y el conjunto de prueba de 400 frases.

Además, se utilizó otro corpus de prueba grabado por teléfono para observar el comportamiento de los tres reconocedores en un ambiente diferente. Esto sirvió como una tercera evaluación donde se reflejó su desempeño en condiciones de un sistema real. Este corpus consta de 400 archivos de voz que contienen cadenas de 2 hasta 9 dígitos pronunciados de manera continua.

En la siguiente sección se muestran los resultados obtenidos en las distintas pruebas efectuadas al evaluar el desempeño de la mejor red neuronal.

3 Evaluación de Resultados

Una vez terminada la fase de desarrollo, evaluamos el desempeño final de los reconocedores usando el conjunto de datos de prueba. Esta fase consiste en evaluar el desempeño de la red neuronal que haya obtenido el mayor porcentaje de reconocimiento en la fase de desarrollo, con un conjunto de datos desconocido para la red, es decir que no haya sido utilizado en ninguna de las dos etapas anteriores.

Las tablas 7-9 muestran, para los tres casos, los porcentajes de error obtenidos por sustituciones, inserciones y eliminaciones, así como el error total a nivel de palabras y frases. Los mejores resultados se obtuvieron en el segundo caso donde algunos fonemas se dividieron en partes obteniendose un total de 139 unidades foneticas dependientes del contexto como nodos de salida de la red neuronal. Por otro lado, el desempeño más bajo se observó en el sistema independiente del contexto.

	%Sub.	% Ins.	% Elim.	% Error (palabras)	% Error (frases)
desarrollo	0.38	0.00	0.48	0.86	3.45
prueba	0.22	0.40	0.49	1.11	6.12
teléfono	3.69	11.55	0.24	15.48	43.25

Tabla 7: Sistema independiente del contexto.

	%Sub.	% Ins.	% Elim.	% Error (palabras)	% Error (frases)
desarrollo	0.14	0.10	0.48	0.72	2.30
prueba	0.31	0.04	0.44	0.80	4.27
teléfono	0.34	2.52	0.10	2.96	10.97

Tabla 8: Sistema dependiente del contexto.

	%Sub.	% Ins.	% Elim.	% Error (palabras)	% Error (frases)
desarrollo	0.19	0.00	0.48	0.67	3.45
prueba	0.31	0.35	0.31	0.98	5.32
teléfono	1.28	4.64	0.24	6.16	23.00

Tabla 9: Sistema dependiente del contexto y uso de categorías generales.

4 Conclusiones

En este artículo se ha hecho una descripción de las distintas fases de desarrollo de un reconocedor de propósito específico para el español mexicano, usando las herramientas del CSLU *Toolkit*. Se obtuvieron tres sistemas los cuales fueron probados con dos conjuntos de datos de prueba y se observó que los sistemas que toman en cuenta el contexto fonético obtuvieron un mayor porcentaje de reconocimiento. De manera general, el desempeño de los tres reconocedores ob-

tenidos disminuyó al ser probados en el corpus de prueba grabado por teléfono. Esto se debe principalmente a ruidos provocados por respiración (del locutor), ruidos ambientales de fondo y ruidos de la línea telefónica. Debido a la susceptibilidad que mostraron los sistemas al ruido aumentó la cantidad de errores principalmente de inserción. Por lo tanto, en el desarrollo de sistemas de reconocimiento de voz es importante considerar estos aspectos y modelarlos adecuadamente etiquetando la base de datos a un mayor nivel de detalle, haciendo uso de diacríticos, con el fin de obtener un mejor nivel de desempeño.

El conjunto de datos de voz recolectado está disponible sin costo alguno a instituciones académicas para propósitos de investigación. Para el grupo, el desarrollo de recursos lingüísticos es parte fundamental para desarrollar futuras investigaciones, es por ello que se ha iniciado la colección de datos de voz en diferentes dominios. Mayor información en la página del grupo Tlatoa: http://info.udlap.mx/~sistemas/tlatoa.

Agradecimientos

La realización de este trabajo no hubiera sido posible sin el apoyo prestado por el *Center for Spoken Language Understanding* del *Oregon Graduate Institute of Science and Technology*.

Referencias

Cole R. et al., *A Survey State of the Art in Human Language Technology*, Cambridge University Press, USA, 1996.

Fanty M., *Overview of the CSLU Toolkit*, Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, USA, 1996.

Goldenthal W., *Statistical Trajectory Models for Phonetic Recognition*, Ph.D. Thesis, Massachusetts Institute of Technology, USA, Agosto, 1994.

Hieronymus J., *ASCII Phonetic Symbols for the World's Languages: Worldbet*, AT&T Bell Laboratories, USA, 1993.

Hosom J., Cole R., Fanty M., Schalkwyk J., Yan Y. and **Wei, W.**, *Training Neural Networks for Speech Recognition*, Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, USA., 1996.

Hosom J. and **Cole R.**, "A Diphone-Based Digit Recognition System using Neural Networks", *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Munich, Abril, 1997.

Hosom J., Cole R., Fanty M. and **Colton D.** *Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding*, Portland, June 6, 1997.

Hosom J., Cole,R., and **Cosi P.**, *Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition*, Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, USA, 1998.

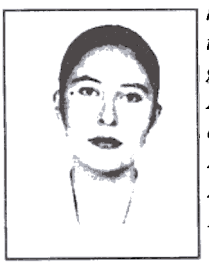
Lander T., *The CSLU Labeling Guide*, Technical Report CSLU-014-96, Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, USA, 1996.

Munive N. and **Vargas A.**, *Reconocedor Fonético de dígitos para el español hablado en México*, Tesis de Licenciatura, Universidad de las Américas-Puebla, México, Diciembre, 1997.

Schalkwyk Johan, Colton Don and **Fanty Mark.**, *The CSLUsh Toolkit for automatic Speech Recognition*. Technical Report Center for Spoken Language Understanding. Oregon Graduate Institute of Science & Technology. Portland, Oregon U.S.A, 1999.

Shultz T., Koll D. and **Waibel A.**, "Japanese LVCSR on the spontaneous scheduling task with Janus-3", *EUROSPEECH*, Rhodes, 1997.

Tapias D., Acero A., Esteve J. and **Torrecilla J.C.** "The VESTEL Telephone Speech Database", *ICSLP*, 1994.



Nora Iliana Munive Mendívil. *Licenciatura en Ingeniería en Sistemas Computacionales Universidad de las Américas, Puebla Actualmente Asistente de Investigación (desde 1997) Grupo de Procesamiento Automático de Voz, Tlatoa Centro de Investigación en Tecnologías de Información y Automatización (CENTIA) Universidad de las Américas, Puebla. Areas de interés: Interacción Humano-Computadora, Reconocimiento de Voz, Diseño de Bases de Datos de Voz.*

Alcira Vargas González. *Grado Académico: Licenciatura en Ingeniería en Sistemas Computacionales Universidad de las Américas, Puebla Actualmente Asistente de Investigación en el Laboratorio de Procesamiento Automático de Voz, Tlatoa que forma parte del Centro de Investigación en Tecnologías de Información y Automatización (CENTIA) en la Universidad de las Américas, Puebla Areas de interés: Procesamiento Automático de Voz, Recuperación de Información e Interacción Humano-Computadora.*

Benjamín Serridge. *Grado Académico: Master en Ciencias Computacionales por el Instituto de Tecnología de Massachusetts (Massachusetts Institute of Technology). Su área de investigación es reconocimiento de voz y diseño de sistemas de lenguaje hablado. Actualmente coordina ALTech-Mexico, una compañía de reconocimiento de voz norteamericana que colabora con la Universidad de las Américas Puebla. Antes, Profesor invitado en la Universidad de las Américas para realizar investigación e impartir cursos de reconocimiento de voz.*

Ofelia Cervantes Villagomez. *Grado Académico: Doctorat Nouveau Regime en Informatique Institut National Polytechnique, Francia, 1988. Diplome d'Etudes Approfondres DEA Ensimag-Institute National Polytechnique, Francia, 1984. Ingeniería en Sistemas Computacionales Universidad de las Américas-Puebla, México, 1981. Actualmente es decana de Asuntos Internacionales y profesor de tiempo completo del Departamento de Ingeniería en Sistemas Computacionales de la UDLA-P.*
Areas de interés: Procesamiento de Señales de Voz, Sintesis de Voz, Bases de Datos Inteligente, Bases de Datos Orientadas a Objetos.

Ingrid Kirschning de Ayala. *Doctorado en Inginería en Systemas de la University of Tokushima, Japón, en marzo de 1998. M.E. en Ciencias de la Computación y Sistemas Inteligentes de la Universidad de Tokushima, Japón en marzo de 1995. Licenciatura en Ingeniería en Sistemas Computacionales de la Universidad de las Américas, Puebla, México (UDLA-P) en 1992. Actualmente es el director del grupo Tlatoa (CENTIA) y profesor de tiempo completo del Departamento de Ingeniería en Sistemas Computacionales de la UDLA-P. Areas de interés: Redes Neuronales Artificiales, Procesamiento Automático de Voz, Recuperación de Información.*