Manifiesto de Digitalización

La siguiente obra es consultable a texto completo; más no se pueden realizar búsquedas de texto en ésta ya que está compuesta por imágenes digitalizadas a partir del original impreso disponible en la biblioteca del área. El siguiente *manifiesto* señala los detalles más resaltantes del mencionado original que inciden en la calidad visual y estructural de este documento electrónico.

Identificación de la obra:

Título:	**Técnicas de aprendizaje artificial aplicadas al problema inverso de la síntesis articulatoria de voz por computadora.**
Autor:	Brito Boadas, José Alejandro
Cota:	QA76.87 B7t

Observaciones sobre el original impreso:

No hay observaciones.

Fecha de digitalización: 10 de marzo de 2009

Todas las páginas fueron digitalizadas en blanco y negro para lograr el mínimo tamaño de archivo posible, siendo la excepción aquéllas en donde el color es crucial para la comprensión de la información.

UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERIA
PROGRAMA DE DOCTORADO EN CIENCIAS APLICADAS
Núcleo Universitario La Hechicera, Edificio B ala Sur, 3er piso
Mérida Estado Mérida. Telefax: 0274- 2403873

ACTA DE EXAMEN DE TESIS DOCTORAL

I. CONSTITUCIÓN DEL JURADO.

En Mérida a las 09:00 am del día Viernes 24 de Noviembre de 2006, se constituyó el Jurado del Examen de Tesis Doctoral del candidato/a: _____

BRITO BOADAS JOSE ALEJANDRO

Cumpliendo con los requerimientos exigidos en el artículo 6 del Reglamento del Doctorado en Ciencias Aplicadas, el jurado quedó conformado de la siguiente manera:

Presidente:
Dr. José Luis Paredes

Tutor:
Dr. Wladimir Rodríguez

Evaluador de la Universidad de Los Andes:

Dra. Elsa Mora

Evaluador de otra Institución:

Dr. Miguel Strefezza

Miembro invitado:

_____

Miembro invitado:

_____

UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERIA
PROGRAMA DE DOCTORADO EN CIENCIAS APLICADAS
Núcleo Universitario La Hechicera, Edificio B ala Sur, 3er piso
Mérida Estado Mérida. Telefax: 0274- 2403873

II. EXAMEN DE TESIS DOCTORAL.

"TÉCNICAS DE APRENDIZAJE ARTIFICIAL APLICADAS AL PROBLEMA INVERSO DE LA SÍNTESIS ARTICULATORIA DE VOZ POR COMPUTADORA"

Calificación (**Menciones posibles**: Aprobado, Aprobado con correcciones formales, Improbado con derecho a un nuevo examen, Improbado sin derecho a un nuevo examen):

APROBADO

Observaciones:

En Mérida a las 12 m del día 24 de Noviembre de 2006 .

El Jurado:

Dr. José Luis Paredes
Presidente
(Universidad de Los Andes)

Dr. Wladimir Rodríguez
Tutor
(Universidad de Los Andes)

Dra. Elsa Mora
Evaluador Interno
(Universidad de Los Andes)

Dr. Miguel Strefezza
Evaluador Externo
(Universidad Simón Bolívar)

Universidad de Los Andes
Consejo de Estudios de Postgrado
Facultad de Ingeniería
Doctorado en Ciencias Aplicadas

Técnicas de Aprendizaje Artificial Aplicadas al Problema Inverso de la Síntesis Articulatoria de Voz por Computadora

Autor
M.Sc. José Alejandro Brito Boadas

Tutor
Dr. Wladimir José Rodríguez Graterol

Investigación presentada ante la Universidad de Los Andes
para optar al grado de Doctor en Ciencias Aplicadas
de la Facultad de Ingeniería

Mérida, 27 de Noviembre de 2006

Resumen

Esta investigación describe el problema inverso de la síntesis articulatoria y la aplicación de técnicas de aprendizaje artificial para su solución, específicamente, redes neuronales y algoritmos genéticos. La síntesis articulatoria recurre a tres modelos: el articulatorio, el acústico, y el de la fuente de excitación. El primero reúne a varios articuladores sobre el plano medial, logrando una reducción del espacio de búsqueda respecto a los modelos estrictamente basados en la función de área. Además, los cambios en la configuración medial se encuentran determinados por contracciones de un conjunto de músculos supraglotales agrupados en el vector articulatorio. Como en la actualidad no existen métricas precisas sobre los efectos e interconexiones de estos músculos, se apela a otra técnica de computación inteligente, el modelado con reglas difusas, para representar la actividad muscular asociada al desplazamiento de la masa lingual. Por su parte, el modelo acústico abarca la propagación de la onda en los tractos supraglotales. Concretamente, la inversión recupera los mejores vectores articulatorios para reproducir las características de frecuencia de un grupo de señales objeto, utilizando los tres modelos. En este sentido, aquí se desarrollan experimentos de inversión con las vocales y las consonantes /m/, /n/, /f/ y /s/, grabadas a locutores venezolanos. Con los fonemas sonoros, una Red con Estados de Eco modela la fuente de excitación. Las señales de entrenamiento de dicha red provienen de un modelo mecánico de dos masas del sistema glotal. Para las fricativas, la excitación es una fuente de turbulencia. Después, mediante un Algoritmo Genético Continuo, se evolucionan poblaciones de configuraciones articulatorias mediales para aproximar las características acústicas de las señales de voz objeto. Los valores de la función objetivo, junto a las evaluaciones subjetivas desarrolladas, verifican positivamente la efectividad de las técnicas inteligentes.

Palabras Clave: Problema inverso del habla, síntesis articulatoria, modelo articulatorio medial, modelo acústico, aprendizaje artificial, redes con estados de eco, lógica difusa, algoritmos genéticos.

Lista de Símbolos

α	Función de análisis acústico de señales
α_C	Función para el cálculo de antiresonancias
α_r	Radio espectral de una ESN
$\delta A(x,t)$	Variación de la función de área respecto a su valor nominal
η	Constante de gas adiabático (1.4)
λ	Coeficiente de conducción calórica del aire (0.55×10^{-3} cal/cm-s-grad)
μ	Viscosidad del aire (1.86×10^{-4} dina-s/cm^2)
ω	Frecuencia angular (rad/s)
ϕ	Sintetizador articulatorio
Ψ	Rango de activación muscular ([0.0, 1.0])
ρ	Densidad atmosférica promedio (1.14×10^{-3} gm/cm^3)
ξ	Calor específico (0.24 cal/gm-grad)
A	Área de un cilindro específico (cm^2)
$A(x,t)$	Función de área
$A_0(x,t)$	Valor nominal de la función de área
A_D	Dominio articulatorio
A_F	Dominio acústico
A_g	Área glotal (cm^2)
AGC	Algoritmo Genético Continuo
AM_i	Actividad muscular en el i-ésimo componente del vector articulatorio
$B(x,y)$	Centro de la masa lingual
b_w	Amortiguamiento por unidad de longitud de las paredes del tracto vocal (1400 gm/s)
C	Capacitancia acústica
c	Velocidad del sonido (cm/s)
$C(x,y)$	Centro del modelo articulatorio
C_1	Resorte de acoplamiento inferior del modelo glotal (cm^2/dina)
C_2	Resorte de acoplamiento superior del modelo glotal (cm^2/dina)
C_c	Resorte de acoplamiento entre masas del modelo glotal (cm^2/dina)
C_w	Capacitancia para el modelado de las pérdidas por vibración de las paredes del tracto vocal
d_i	Nivel de activación de una regla difusa
DR	Reservorio Dinámico de una Red con Estados de Eco
$e(x,t)$	Voltaje en una línea de transmisión

E_i	Puntos de la frontera posterior-superior del modelo articulatorio
EAL	Espacio Acústico Libre
ESN	Red con Estados de Eco
f	Frecuencia cíclica (Hz)
$F0$	Frecuencia fundamental de vibración glotal (Hz)
f_1	Término de la función objetivo del AGC que cuantifica la distancia acústica entre señales
F_1, F_2	Locutores femeninos del corpus de señales objeto
f_2	Término de la función objetivo del AGC que introduce el criterio de mínima actividad muscular
f_x	Función de activación sigmoidal de las unidades internas de la ESN
f_y	Función de activación sigmoidal de las unidades de salida de la ESN
Fn	N-ésima fecuencia formante (Hz)
G	Resistencia que modela las pérdidas por disipación térmica en el tracto vocal
GGa	Geniogloso Anterior
GGm	Geniogloso Medio
GGp	Geniogloso Posterior
h	Dimensión característica de una constricción
$H(\omega)$	Función de transferencia
HG	Hiogloso
$i(x,t)$	Corriente en una línea de transmisión
I_i	Puntos de la frontera anterior-inferior del modelo articulatorio
$J(x,y)$	Unión temporomandibular
k_w	Dureza por unidad de longitud de las paredes del tracto vocal (3000 dinas/cm)
L	Inductancia acústica
l	Longitud de un cilindro específico (cm)
L_w	Inductancia para el modelado de las pérdidas por vibración de las paredes del tracto vocal
LoI	Músculos intrínsecos de la lengua para descenso del ápice
M_1	Masa inferior del modelo glotal (gm/cm)
M_1, M_2	Locutores masculinos del corpus de señales objeto
M_2	Masa superior del modelo glotal (gm/cm)
$M_i(\omega)$	Matriz ABCD de una red eléctrica T
M_r	Probabilidad de mutación del AGC
m_w	Masa por unidad de longitud de las paredes del tracto vocal (1.5 gm)
M_{ij}	Conjuntos difusos de salida
MA	Masetero
MC	Constrictor Medio de la Faringe
MH	Milohioideo
n_t	Cantidad de pasos de estabilización del DR de una ESN durante el entrenamiento
N_{POB}	Tamaño de la población del AGC
OO	Orbicular de los labios
$p(t)$	Vector articulatorio
$P(x)$	Presión del aire (dinas/cm^2)
$P(z)$	Polinomio en la variable compleja z
P_g	Presión de aire en la glotis (dinas/cm^2)
P_s	Presión de aire subglotal (dinas/cm^2)
PL	Predicción Lineal
R	Resistencia que modela las pérdidas por fricción viscosa en el tracto vocal

R_e	Número de Reynolds
R_w	Resistencia para el modelado de las pérdidas por vibración de las paredes del tracto vocal
R_{ec}	Número crítico de Reynolds (2700)
RI	Risorio
S	Longitud de la circunferencia de un cilindro específico
S_v	Señal objeto
SG	Estilogloso
T_s	Período de muestreo del Modelo Acústico (ms)
$turbg$	Ganancia de turbulencia (20×10^{-6})
$U(x)$	Velocidad del volumen de aire (cm^3/s)
U_g	Velocidad del volumen de aire en la glotis (cm^3/s)
$u_i(n)$	Nivel de activación de las unidades de entrada de la ESN
UpI	Músculos intrínsecos de la lengua para ascenso del ápice
V	Segmento de máximo descenso del paladar blando (magnitud de 1.0 cm)
v	Velocidad lineal del flujo de aire (1000 cm^3/s)
W_b	Pesos de sinapsis de retroalimentación de la ESN
w_i	Peso asociado a una regla difusa
W_u	Pesos de sinapsis de entrada de la ESN
W_x	Pesos de sinapsis internas de la ESN
W_y	Pesos de sinapsis de salida de la ESN
$x_i(n)$	Nivel de activación de las unidades internas de la ESN
$y_i(n)$	Nivel de activación de las unidades de salida de la ESN
Z_1,Z_2	Impedancias de una red eléctrica T
Z_B	Impedancia del tracto oral
Z_N	Impedancia del tracto nasal
/a/	Vocal baja-central
/b/	Consonante oclusiva, bilabial, sonora
/e/	Vocal media-anterior
/f/	Consonante fricativa, labiodental, sorda
/i/	Vocal alta-anterior
/m/	Consonante nasal, bilabial, sonora
/n/	Consonante nasal, alveolar, sonora
/o/	Vocal media-posterior
/p/	Consonante oclusiva, bilabial, sorda
/s/	Consonante fricativa, alveolar, sorda
/u/	Vocal alta-posterior

Índice general

Índice de figuras

Índice de cuadros

1

El Problema Inverso de la Síntesis Articulatoria

1.1. Introducción

Las Tecnologías del Habla constituyen un área de investigación interdisciplinaria que involucra conocimientos de Mecánica, Matemáticas, Computación, Fonética, Procesamiento de Señales, entre otros. En la persistente búsqueda de reproducir (acaso superar) en lo artificial las propiedades de lo natural, y en específico, de la inteligencia, muchos investigadores han perseguido la meta de construir una máquina con capacidades de pronunciación inteligible. Existen antiquísimas referencias sobre este propósito: tres siglos A.C., Herón de Alejandría disertaba acerca de autómatas parlantes [55]. En general, el problema consiste en producir automáticamente mensajes orales, a partir de una representación simbólica, frecuentemente, un texto. Con el surgir de las computadoras digitales la experimentación ha progresado enormemente, distinguiéndose tres tendencias en la síntesis: por concatenación, por formantes y articulatoria. Esta última ha sido señalada como la posible respuesta definitiva para la síntesis de alta calidad, a largo plazo [33]. Empero, actualmente la calidad en las emisiones sintéticas de los métodos articulatorios resulta inferior a la de los otros. La dificultad con la síntesis articulatoria reside en el modelado de un sistema biomecánico y acústico altamente complejo: el aparato fonador humano. Adicionalmente, indaga en la producción y comportamiento de la energía acústica en las cavidades del sistema. La recolección de los datos pertinentes resulta muy complicada, por lo que obligatoriamente se trabaja con información incompleta. Para solventar esta carencia informativa, la presente investigación aplica métodos de inversión de la síntesis, con el fin de recuperar datos articulatorios a partir de señales verbales objeto. Las secciones siguientes completan la descripción del problema, incluyendo los enfoques e intereses propios del estudio.

1.2. Planteamiento del Problema

La síntesis articulatoria de voz transforma un vector $p(t)$ de parámetros anatómicos o fisiológicos en una señal verbal S_v con características acústicas predefinidas [54]. Por ejemplo, $p(t)$ puede incluir la posición del centro de la masa lingual y del hioides, la protrusión y abertura de los labios, el área del paso velofaríngeo, entre otros parámetros articulatorios. Por su parte, un sintetizador articulatorio $\phi : A_D \rightarrow F_D$ mapea el dominio articulatorio A_D (sobre el cual se define $p(t)$) en el dominio acústico F_D (al cual pertenecen las propiedades de frecuencia de S_v). Una función α retorna el vector columna con las propiedades acústicas de una señal. Ahora, mediante estas definiciones, el problema inverso de la síntesis articulatoria, también llamado *del habla*, puede expresarse como un problema de optimización:

$$min_{p(t)\in A_D} \left\{ W^T \left| \alpha(\phi(p(t))) - \alpha(S_v) \right| \right\} \tag{1.1}$$

donde W es un vector columna que refleja la importancia relativa de cada componente de error. En palabras, la meta consiste en recuperar los parámetros que, una vez introducidos al sintetizador, le permiten generar una señal lo más similar posible, acústicamente, a la señal objeto S_v predefinida. Idealmente, los parámetros articulatorios deben derivarse del análisis de la propia señal objeto. Por tal razón, el problema también recibe el nombre de *mapeo acústico → articulatorio.* Evidentemente, se trata del complemento del problema de la síntesis articulatoria de voz o *mapeo hacia adelante,* representado por ϕ. La síntesis articulatoria recurre al modelado del sistema de producción de voz que incluye, entre otros componentes, los articuladores presentados en la Figura 1.1.

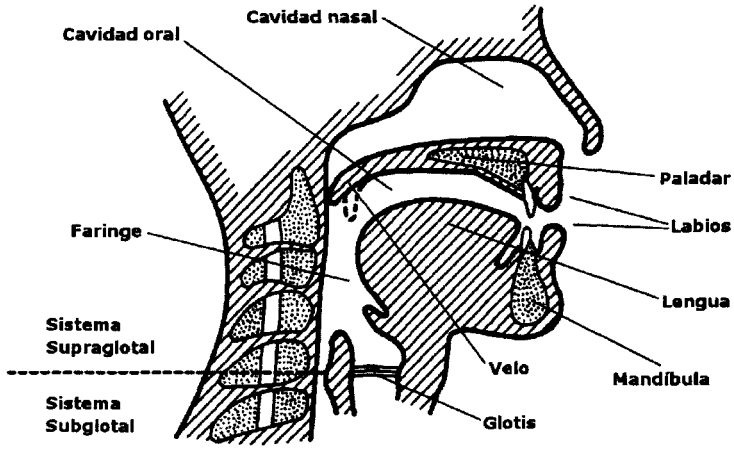

Figura 1.1: *Aparato Fonador (parcial) [27].*

1.3. MODELOS DE LA SÍNTESIS

Básicamente, ϕ modela la dinámica de los articuladores, el flujo de aire resultante por la acción de los articuladores sobre la señal de excitación, y las fuentes de energía. Con este propósito, ϕ integra tres modelos: el *modelo articulatorio,* el *modelo acústico,* y el *modelo de la fuente de excitación.* Un modelo articulatorio representa los componentes esenciales de la producción del habla, y su propósito principal es el cómputo de la función de área $A(x,t)$, la cual refleja la variación en el área de la sección transversal del tubo acústico cuyas fronteras se ubican en la glotis y en la boca. El área se evalúa respecto a la distancia (variable x) y al tiempo (variable t). Para los sonidos nasales también debe considerarse el área del tracto nasal[1]. En resumen, el modelo articulatorio describe la geometría de los tractos supraglotales.

Por su parte, el modelo acústico especifica las transformaciones entre $A(x,t)$ y F_D, determinadas por la propagación de la onda sonora en los tractos. De acuerdo con la teoría acústica de la producción del habla [21, 64], las señales de voz se consideran salidas de un filtro caracterizado por $A(x,t)$ y excitado por alguna fuente de energía, como ilustra la Figura 1.2.

Por último, el modelo de la fuente de excitación representa las fuentes de energía y su posible interacción con los tractos. Para los fonemas sonoros, como las vocales, la energía corresponde a una modulación del flujo subglotal por parte de las cuerdas vocales, mientras que para otros

[1]Sin embargo, a diferencia de lo que ocurre con $A(x,t)$, el área del tracto nasal no experimenta cambios significativos durante la pronunciación. Solamente su sección inicial, vinculada al área del paso velofaríngeo, puede variar sus dimensiones con el tiempo.

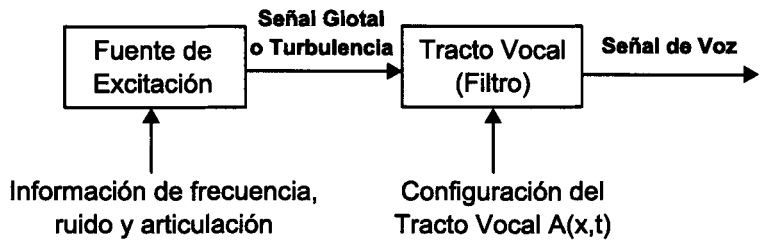

Figura 1.2: *Modelo Fuente-Filtro para la producción de la voz.*

fonemas, como las fricativas, la señal de excitación es una corriente de turbulencia formada en alguna constricción supraglotal. La fuente para los fonemas oclusivos, por el contrario, resulta de una liberación abrupta de energía retenida por la oclusión, derivándose así una energía de excitación impulsiva. En consecuencia, el alcance de este modelo depende de la clase de fonemas abordados por la inversión. Pero por sí mismo, el tema resulta bastante complicado. Por ejemplo, Sinder dedica toda su tesis doctoral al desarrollo de un modelo exclusivo para la generación de consonantes fricativas [69].

1.4. Inversión del Vector Articulatorio

Una vez definidos los modelos de ϕ, se procede a determinar con alguna precisión la relación entre $p(t)$ y S_v. Sin embargo, como el mapeo entre el dominio acústico y el articulatorio resulta no lineal y muchos-a-uno [10–12,61,62,67], la definición y alcance de soluciones aceptables al problema inverso no constituyen tareas triviales. En general, la evaluación de una solución potencial sigue algún tipo de relación sobre F_D, al estilo de la Ecuación 1.1. Por ejemplo, sobre F_D pueden definirse los formantes o la densidad de energía espectral, de modo que la diferencia entre estas métricas aplicadas a la señal objeto y a la señal sintética represente el error del proceso de inversión. Pero algunos fonemas, como los oclusivos, constan de una serie de eventos acústicos complejos, más difíciles de cuantificar.

El carácter no lineal proviene de la dinámica subyacente al aparato fonador, sistema que varía en el tiempo, y cuyos componentes exhiben un alto nivel de acoplamiento indirecto a través de músculos que los conectan. Adicionalmente, la forma y frecuencia de la señal de excitación glotal constituyen funciones no lineales [62], con evidencia de caos en la vibración del tejido glotal [84, 95]. Por otra parte, existe un acoplamiento natural entre la señal fuente y el filtro, manifestado en la influencia que tiene la masa acústica en las cavidades supraglotales sobre el grado de inclinación del pulso glotal [78]. De este modo, mínimas variaciones en la trayectoria de los articuladores alteran la salida significativamente.

Por otro lado, el mapeo muchos-a-uno implica que configuraciones articulatorias muy diferentes pueden producir sonidos bastante parecidos [15]. Esta idea se ilustra en la Figura 1.3. Las articulaciones compensatorias de los ventrílocuos y de algunas aves, principalmente las pertenecientes a la familia de los psitácidos, como los loros, representan instancias notables de esta condición [22]. Las articulaciones compensatorias ajustan las dimensiones de las cavidades resonantes para reproducir las características acústicas de otra configuración. Puede proporcionarse un ejemplo más formal de este obstáculo de la inversión articulatoria. La variación de presión $p(x)$ a lo largo de un tubo acústico de área uniforme ($A(x,t) = A =$ constante) como el de la Figura 1.4 viene dada por la Ecuación 1.2 [78], donde f denota la frecuencia en Hz, y c es la velocidad del sonido (3.53×10^4 cm/s a 37°C) [16].

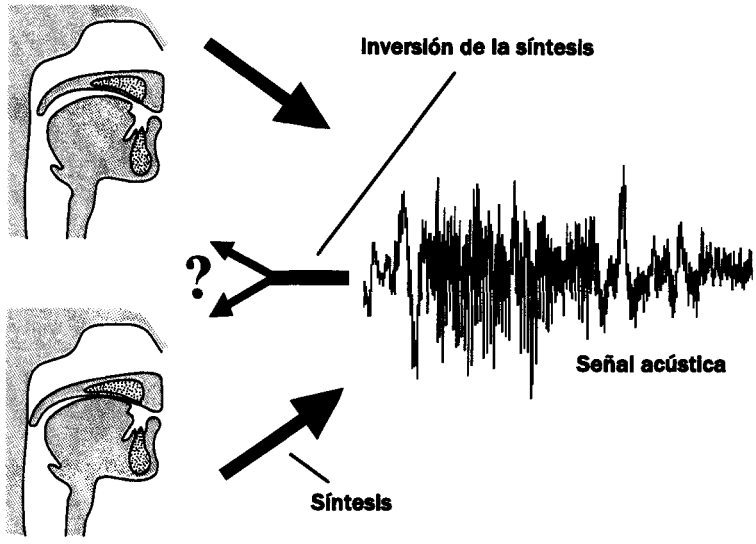

Figura 1.3: *Problema de Inversión [15]. La figura sólo ejemplifica; de ningún modo las configuraciones se relacionan acústicamente con la señal dibujada.*

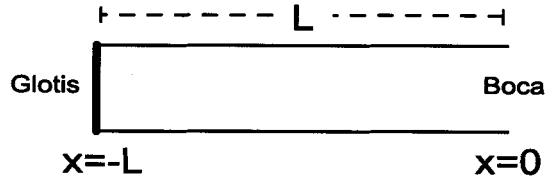

Figura 1.4: *Tubo acústico de área uniforme.*

$$\frac{d^2p}{dx^2} + \left(\frac{2\pi f}{c}\right)^2 p = 0 \tag{1.2}$$

La Figura 1.4 evidencia que la impedancia acústica en la boca es cero, y la glotal, infinita. La solución, bajo estas condiciones de frontera, es

$$p(x) = P_m \sin \frac{2\pi f}{c} x \tag{1.3}$$

donde P_m es el pico de presión sonora. Por otro lado, la Ecuación 1.4 formaliza la relación de la presión con la velocidad del volumen de aire [78].

$$\frac{dp}{dx} = -\frac{j2\pi f\rho}{A} U \tag{1.4}$$

Luego, a partir de $p(x)$ y la Ecuación 1.4, la velocidad del volumen viene dada por

$$U(x) = jP_m \frac{A}{\rho c} \cos \frac{2\pi f}{c} x \tag{1.5}$$

con ρ igual a la densidad atmosférica promedio (1.14×10^{-3} gm/cm^3 a 37°C) [78]. Como $U(-L) = 0$, las resonancias Fn del tubo acústico son

$$Fn = \frac{2n-1}{4}\frac{c}{L} \tag{1.6}$$

con $n = 1, 2, 3...$ Según la Ecuación 1.6, la función de área no interviene en la ubicación de las resonancias del tubo en la Figura 1.4. Entonces, bajo las condiciones anteriores, configuraciones de área disímil tendrían resonancias idénticas si ajustan su longitud a un valor común. Como segunda consecuencia, la inversión articulatoria debe considerar también las longitudes de los tractos, principalmente si el corpus de inversión incorpora señales de diversos locutores. La inversión debe discernir, por ejemplo, entre señales de locutores masculinos y femeninos, por cuanto la longitud promedio del tracto vocal masculino es de 16.9 cm, y del femenino, 14.1 cm [19].

Adicionalmente, de la familia de soluciones al problema, con frecuencia interesan sólo aquellas consistentes con las descripciones provistas por la fonética articulatoria, lo cual implica que la inversión debe dirigirse hacia el grupo de configuraciones anatómicamente lícitas. Por otra parte, la imprecisión en cuanto a la naturaleza y cantidad de variables articulatorias y acústicas introduce otro inconveniente: prácticamente cada investigación define su propio conjunto de variables, sin que exista un acuerdo definitivo sobre el número y significado de las mismas. En consecuencia, al referirse a la recuperación de parámetros articulatorios a partir de una señal acústica, la noción de lo que constituye un parámetro depende del modelo articulatorio de la investigación, explicado en el Capítulo 3.

1.5. APLICACIONES DE LA INVERSIÓN ARTICULATORIA

Entre otras, la solución al problema inverso resulta de interés para las siguientes aplicaciones:

1. Compresión de la señal de voz, porque en general las dimensiones de la señal superan notablemente a las del vector articulatorio. En este sentido, dicho vector puede utilizarse como representación de la señal de voz para el almacenamiento y transmisión de la misma.

2. Recolección y análisis de datos sobre los procesos de la fonación a bajo costo, y de forma no invasiva, complementando los datos provenientes de la electropalatografía y la resonancia magnética, técnicas que resultan costosas (en tiempo y dinero) y sensibles al ruido.

3. Reconocimiento de voz, por medio de la transición al dominio articulatorio A_D, donde las señales pueden caracterizarse con menos parámetros. Recuérdese que el reconocimiento, en última instancia, consiste en transformaciones entre dominios. Mediante la inversión, por ejemplo, las muestras de una vocal con 25 ms de duración, discretizada a 10 kHz, pueden transformarse en sólo 12 parámetros de un vector articulatorio.

4. Recuperación de los mejores parámetros para síntesis de señales de voz de alta calidad. En el largo plazo, a medida que las bases teórico-prácticas de la síntesis articulatoria se consoliden, y la capacidad de cómputo aumente, este tipo de síntesis debería convertirse en la más idónea para los sistemas de conversión de texto a voz, por su carácter parametrizable y su capacidad para recrear diversos eventos acústicos [33].

5. Copia de un locutor, aplicación interesante para la industria del entretenimiento, entre otras. Al aproximarse al vector articulatorio correcto, la inversión copia la señal objeto en alguna

medida. No obstante, la única forma de efectuar una copia *perfecta* es que todos los modelos empleados puedan reproducir las propiedades articulatorias y acústicas del locutor específico invertido.

1.6. ANTECEDENTES

La aproximación tradicional para el problema de inversión recurre a los libros de código con parámetros articulatorios [37, 67]. Un libro de código es una tabla extensa, frecuentemente con más de cien mil entradas que representan configuraciones diversas del tracto vocal, acompañadas de su respectiva salida acústica, sintetizada. Luego, la señal objeto se analiza por tramas, y se busca en el libro aquella configuración con la salida más cercana a la señal objeto. La cercanía viene dada por una función de costo basada en el análisis acústico, y la búsqueda apela a algoritmos de programación dinámica. Sin embargo, el enfoque resulta computacionalmente costoso, porque la síntesis articulatoria exige la solución de muchas ecuaciones simultáneas, y deben sintetizarse las señales asociadas a todas las configuraciones del libro. Y en su mayoría el libro de código alojará configuraciones inútiles para una señal objeto específica.

Como alternativa, puede intentarse la reducción del espacio acústico a regiones articulatorias mediante técnicas de clustering [59]. Para ello, se mapea F_D en A_D mediante redes perceptrónicas multicapa. Posteriormente, la búsqueda procede sobre las regiones, y no sobre los vectores acústicos completos. Desafortunadamente, el clustering sacrifica información acústica que puede resultar indispensable para la inversión de futuras señales objeto. También se han empleado las redes neuronales para el mapeo directo entre el dominio acústico y el articulatorio, sin clustering. Por ejemplo, en [53] se utilizan datos de radiografías para que la señal aprenda la correspondencia entre configuraciones del tracto vocal y pronunciaciones. Empero, la red posee limitaciones severas en cuanto a la cantidad de fonemas que puede aprender. Continuando con los modelos conexionistas, en [60] han entrenado un perceptrón multicapa para mapear entre el espectro de energía de vocales y consonantes y los parámetros de un modelo del tracto vocal, definido por un número fijo de secciones transversales de cilindros acústicos. En un enfoque relacionado, se ha utilizado una red modificada para considerar las restricciones derivadas de las propiedades cinemáticas del sistema de producción de voz [4]. Con ello se ha logrado generalizar el patrón de movimientos, interpolando nuevas trayectorias a partir de las aprendidas.

Otra opción consiste en apelar a los algoritmos genéticos. McGowan [45] agrupó los tres primeros formantes de la señal en el vector acústico, e introdujo restricciones al modelo articulatorio, concretamente, en las variables del tracto vocal. Dichas variables describen los grados y ubicación de las restricciones en el modelo anatómico. El estudio aborda las relaciones entre la articulación y la percepción sobre la base de *tareas*, a partir de una descripción dinámica de entradas a un sintetizador [32, 65]. Luego, se recurre a los algoritmos genéticos porque dichas restricciones resultaron muy difíciles de abordar con técnicas de descenso por gradiente. Sólo logró recuperarse parcialmente la trayectoria articulatoria de dos secuencias Vocal-Consonante-Vocal, con la desventaja de pérdida de precisión por la codificación binaria.

En general, el problema de inversión resulta muy difícil. Existen algunas investigaciones que intentan enfoques alternos, con resultados regulares. Por ejemplo, Blackburn y Young [7] experimentan con secuencias fonéticas alineadas en el tiempo, mientras que Yehia e Itakura [94] y Ouni y Laprie [51] también experimentan con restricciones, si bien en modelos articulatorios particulares, restringidos. Por su parte, Dusan y Deng desarrollaron métodos analíticos para recuperar las configuraciones del tracto vocal [20]. Estudios más recientes emplean puntos de control medidos experimentalmente en un grupo de locutores, y la inversión procede tratando de minimizar, mediante

aproximación cuadrática, la distancia entre el modelo articulatorio y dichos puntos [38, 39, 75–77].

No obstante, la presente investigación se diferencia de las anteriores en las señales y técnicas inteligentes empleadas para el modelado del aparato fonador y para la inversión con base en datos acústicos, como se describe en los próximos capítulos. Allí reside el carácter novel de este trabajo, confiriéndole el estatus de primigenio.

1.7. TÉCNICAS DE APRENDIZAJE ARTIFICIAL

Para afrontar las dificultades de la inversión articulatoria, esta investigación recurre a algunas técnicas de Aprendizaje Artificial. Con detalle, se rechaza en lo posible el conocimiento a priori sobre la articulación de S_v, y se intenta, mediante técnicas concretas, que la máquina *aprenda* cuál es la configuración articulatoria correspondiente. En general, un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y función de evaluación P, si su desempeño en las tareas T, según la métrica P, mejora con la experiencia E [48]. La Ecuación 1.1 define la métrica P, por lo que el error y el aprendizaje mantienen una relación inversamente proporcional. A su vez, la información de error proveniente del análisis de las configuraciones articulatorias representa la experiencia E, mientras que T corresponde a la síntesis articulatoria de la señal objeto S_v. En este sentido, la investigación recurre a los Algoritmos Genéticos Continuos [24] para la recuperación de la información articulatoria, principalmente, por la aptitud de dicha técnica para codificar las configuraciones del modelo articulatorio, y para efectuar una búsqueda estocástica sobre el espacio articulatorio.

No obstante, la recuperación del vector articulatorio constituye sólo una parte de toda la inversión. Existen otros problemas, como la construcción de los modelos de la síntesis, en los cuales la introducción de las técnicas de aprendizaje o de inteligencia computacional pueden contribuir. Específicamente, aquí se recurre a una red neuronal recurrente y a un sistema de inferencia difusa para el modelado de la excitación glotal y de la dinámica de la masa lingual, respectivamente.

1.8. TIPO DE INVERSIÓN

La validación acústica de la inversión exige la implementación computacional de ϕ. En consecuencia, todas las señales de la investigación, incluyendo S_v, son discretas. Además, las señales objeto en el corpus se preprocesan manualmente, con la finalidad de remover los silencios y las transiciones hacia los fonemas vecinos, cuando aplique. De esta forma, las señales del corpus exhiben características de frecuencia relativamente estáticas, y la inversión de cualquiera de estas señales requeriría un único vector articulatorio. Así, no resultaría necesaria la dependencia del tiempo en el vector articulatorio y en la función de área. En todo caso, la síntesis de secuencias (por ejemplo, Consonante-Vocal), útiles en la evaluación subjetiva, sí requiere la dependencia temporal, por lo que en el resto del trabajo el vector articulatorio continuará denotándose $p(t)$, y el área, $A(x,t)$. Con la señal objeto, por el contrario, no hay necesidad de hacer explícita la dependencia, por lo que se refiere simplemente como S_v. En conclusión, la inversión clasifica como **estática**, y por consiguiente se recuperan configuraciones articulatorias, y no trayectorias.

1.9. APORTES DE LA INVESTIGACIÓN

En la búsqueda de soluciones al problema inverso se han desarrollado numerosos métodos. Las principales contribuciones de esta investigación son:

1. Primera investigación que integra diversas técnicas de aprendizaje artificial en la solución del problema inverso.
2. Modela la excitación glotal cuasi-periódica usando Redes con Estados de Eco, representa parte de la actividad muscular supraglotal con un sistema de inferencia difuso fácilmente expansible, y emplea Algoritmos Genéticos Continuos para recuperar las configuraciones sobre el plano medial. Tales técnicas no se habían aplicado anteriormente a la inversión articulatoria.
3. Las métricas que guían la búsqueda del espacio articulatorio se basan en el análisis acústico, y son exclusivas de este trabajo.
4. El vector articulatorio posee un significado fisiológico y considera los descubrimientos recientes.
5. Se extiende el modelo articulatorio clásico de Mermelstein para una mejor representación de la zona del paladar blando.
6. El modelo acústico incorpora fuentes de excitación glotal y de turbulencia.
7. El corpus de señales objeto incluye grabaciones a locutores masculinos y femeninos, por lo que antes de la inversión se ajustan las dimensiones del modelo articulatorio, dependiendo del género del locutor.
8. Se invierten varios fonemas con el mismo núcleo de técnicas. Representa un primer paso hacia la consolidación de técnicas suficientemente generales para abordar cualquier fonema.
9. La investigación proporciona evaluaciones objetivas, subjetivas y también muestra las mejores configuraciones articulatorias recuperadas.

Desde el punto de vista práctico, el estudio aporta la implementación en MATLAB de varias rutinas que pueden apoyar o constituir el punto de partida de futuras investigaciones. Específicamente, se ha codificado:

- El modelo articulatorio con dimensiones escalables y un sistema de inferencia difusa para el modelado de la actividad de los músculos extrínsecos de la lengua. Incluye también procedimientos para el cálculo de mallas con un número variable de cilindros en la aproximación de los tractos supraglotales.
- El modelo acústico tipo Maeda, con pérdidas de energía, e integrable con múltiples fuentes de excitación. Comprende los tractos faríngeo, oral y nasal, junto al modelo de radiación de Flanagan.
- Red con Estados de Eco como modelo de fuente de excitación glotal, y modelo acústico de turbulencia.
- Sintetizador articulatorio configurable por tramas, con interpolación automática de funciones de área y longitud.
- Rutinas para el cómputo de la función de transferencia de vocales, nasales y fricativas.
- Sistema masa-resorte para el modelado de los tejidos glotales.
- Algoritmo Genético Continuo para la inversión de las configuraciones articulatorias.
- Múltiples rutinas de apoyo, por ejemplo, para el entrenamiento de la red neuronal, y para la recopilación de resultados y producción de gráficos.

1.10. Organización de la Tesis

El resto del documento se encuentra organizado de la forma siguiente:

- El **Capítulo 2** aborda la red neuronal recurrente (Red con Estados de Eco [30]) utilizada para el aprendizaje supervisado de señales de excitación glotal. Las señales de entrenamiento se derivan de un modelo mecánico de dos masas del tejido glotal [28].
- El **Capítulo 3** contiene la elaboración completa del modelo articulatorio. Además, incluye un sistema de inferencia difusa para el modelado de la dinámica de la masa lingual. Específicamente, la actividad de algunos músculos extrínsecos de la lengua se modela mediante variables difusas, y las relaciones entre contracción y efecto sobre la configuración medial se aproximan con un sistema de inferencia difusa estilo Takagi-Sugeno-Kang [81, 82].
- El **Capítulo 4** describe el modelo acústico tipo Maeda [42] utilizado durante la validación acústica de la inversión, y en el cálculo de las funciones de transferencia requeridas para determinar las propiedades acústicas de las configuraciones mediales.
- El **Capítulo 5** aborda la inversión del vector articulatorio mediante algoritmos genéticos continuos [24], cuyos cromosomas reales admiten una representación más directa de A_D. La definición de la función objetivo depende de la clase de señal a invertir, pero en general, emplea sólo información sobre la distribución de la energía de las señales en el dominio de la frecuencia, aunada a un criterio de actividad muscular que favorezca las configuraciones con menor gasto energético. Para los experimentos de inversión se conforma un corpus con las señales objeto de las cuales se recuperará el vector articulatorio. Se emplean grabaciones de varios locutores, para confirmar que los métodos de aprendizaje funcionen con diversos datos. Además, las señales objeto se restringen a las vocales, las consonantes nasales /m/ y /n/, y las fricativas /f/ y /s/. Tradicionalmente, las consonantes han resultado los sonidos más difíciles para los sintetizadores articulatorios [16, 69].
- El **Capítulo 6**, finalmente, engloba las conclusiones y recomendaciones de la investigación.

2

Aprendizaje de la Excitación Glotal

2.1. Introducción

La producción de energía en la glotis y en el tracto vocal siempre involucra una modulación del flujo de aire subglotal, en el centro o proximidad de alguna constricción. Así, la excitación para las consonantes fricativas corresponde a una señal de turbulencia o ruido derivada de la colisión de un volumen de aire desplazándose a alta velocidad contra un obstáculo, los incisivos, por ejemplo, ubicado algunos milímetros más abajo de la zona de constricción. Para las oclusivas, la excitación adquiere una forma impulsiva, por la liberación de energía acumulada. Por su parte, la energía para los fonemas sonoros proviene de la modulación, en la glotis, del flujo subglotal. Debido al movimiento lateral aproximadamente periódico de las cuerdas vocales, la excitación glotal también resulta cuasi-periódica. Esta variabilidad en la señal glotal contribuye con el timbre *natural* de la señal radiada. Por el contrario, las señales obtenidas con sintetizadores articulatorios, utilizando una señal de excitación perfectamente periódica, poseen un timbre metálico, antinatural [16]. En este sentido, el resto del capítulo se concentra en el modelado de esta variabilidad en la señal glotal, lo cual también amerita una revisión de la naturaleza y formación de la señal. Se parte del modelo mecánico clásico de las cuerdas vocales [28] para arribar al modelado de la señal con una red neuronal recurrente. La discusión de las fuentes aperiódicas se posterga hasta el Capítulo 4, porque éstas se basan en modelos netamente acústicos.

2.2. Naturaleza de la Excitación Glotal

Esta sección se concentra en la definición del modelo mecánico de las cuerdas vocales, a partir de revisiones de la literatura. El modelo tiene como objetivo principal la aproximación de la dinámica de las cuerdas vocales de locutores prototipo, de género masculino y femenino. Las cuerdas vocales son dos masas de tejido, ligamentos y músculos, conectadas anteriormente al cartílago tiroides, y posteriormente a los dos cartílagos aritenoides. Los aritenoides poseen cierto grado de movilidad, lo que permite variar las dimensiones del espacio entre ambas masas. Dicho espacio recibe el nombre de *glotis*, y su área, vista desde arriba, se denomina *área glotal* A_g. La Figura 2.1 exhibe tres configuraciones importantes de la glotis. En el caso de la Figura 2.1(a) las cuerdas vocales se encuentran relajadas y apartadas, y el aire transita entre ellas sin obstrucción. No obstante, al aproximarse los tejidos, surgen dos posibilidades. La primera, mostrada en la Figura 2.1(b), se emplea durante la deglución, en la cual el cierre de las cuerdas vocales obstruye totalmente el flujo de aire. En la otra opción, pertinente a la fonación, un contacto más laxo permite que el aire pase ejerciendo presión sobre los tejidos, con algo de oposición, como ilustra la Figura 2.1(c). El encuentro del flujo con esta

constricción ocasiona la vibración de los tejidos, y la perturbación se transmite a las moléculas de aire. Además del tamaño de la glotis, los músculos y cartílagos de la zona controlan la tensión del tejido, parámetro vinculado a la frecuencia de las vibraciones o *frecuencia fundamental F0*.

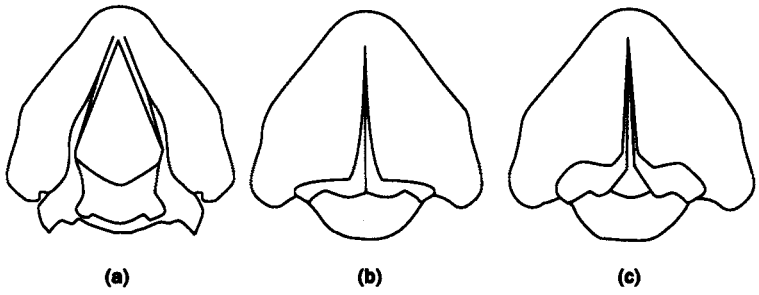

Figura 2.1: *Configuraciones de la glotis [5].*

La Figura 2.2 revela la dinámica aproximada de las cuerdas vocales, correspondiente a la configuración de la Figura 2.1(c), única que origina la modulación del flujo de aire subglotal. El ciclo inicia en la Figura 2.2(a) cuando el aire resultante de la contracción pulmonar ya ha logrado separar por completo las cuerdas vocales. La flecha en la Figura 2.2(a) sigue el sentido ascendente, desde la tráquea hacia las cavidades supraglotales. Debido al principio de Bernoulli, el incremento en la velocidad del volumen de aire ocasiona una caída en la presión sobre el tejido, por lo cual los tejidos inferiores, de mayor volumen, entran en contacto anticipadamente, como muestra la Figura 2.2(b). Por efecto de la conexión elástica entre las masas inferiores y las superiores, y por la mencionada caída de presión, la ranura glotal prosigue su cierre en forma ascendente, estado que se presenta en la Figura 2.2(c). No obstante, apenas la glotis se cierra, en la configuración de la Figura 2.2(b), comienza a crecer la presión subglotal, que forzará una nueva separación de los tejidos inferiores, ilustrada en la Figura 2.2(d). Por el acoplamiento entre tejidos, la separación de las porciones inferiores induce la separación de los superiores, retornando así al estado inicial, etapas reflejadas en las Figuras 2.2(e) y 2.2(f). Obsérvese que la trayectoria de los tejidos superiores exhibe una diferencia de fase respecto al movimiento de los inferiores. Y como estos tejidos elásticos se encuentran conectados, los desplazamientos de ambas porciones se afectan mutuamente. Y dicho vínculo, aunado a una presión subglotal relativamente regular, facilita la repetición del proceso. Claramente, A_g varía entre los extremos representados por la Figura 2.2(a) y la Figura 2.2(f), afectando el volumen de aire a través de la glotis, U_g, durante el período glotal. La Ecuación 2.1 aproxima esta relación, donde P_s es la presión subglotal [78].

$$P_s = \frac{\rho U_g^2}{2A_g^2} \tag{2.1}$$

Mecánicamente, y asumiendo una simetría perfecta en la forma de las cuerdas vocales, los tejidos superiores e inferiores de la glotis pueden modelarse mediante un sistema de masas y resortes como el de la Figura 2.3. El resorte C_c modela la relación de acoplamiento entre los tejidos inferiores y superiores, cuyas masas se encuentran denotadas por M_1 y M_2, respectivamente, con unidades de gm/cm [28]. Por su parte, los resortes C_1 y C_2 representan la conexión de los cartílagos de soporte glotal con el tejido inferior y el superior, respectivamente. Todos los coeficientes de elasticidad del modelo se expresan en cm^2/dina. Existen otros modelos diferenciados principalmente por el grado

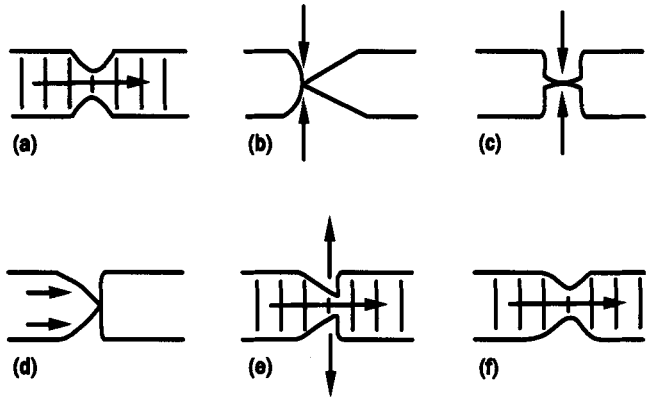

Figura 2.2: *Dinámica glotal [57].*

de distribución de la masa de todo el sistema glotal [80,92]. Empero, para el propósito de aproximar el área glotal, el modelo de dos masas resulta suficiente y de bajo costo computacional.

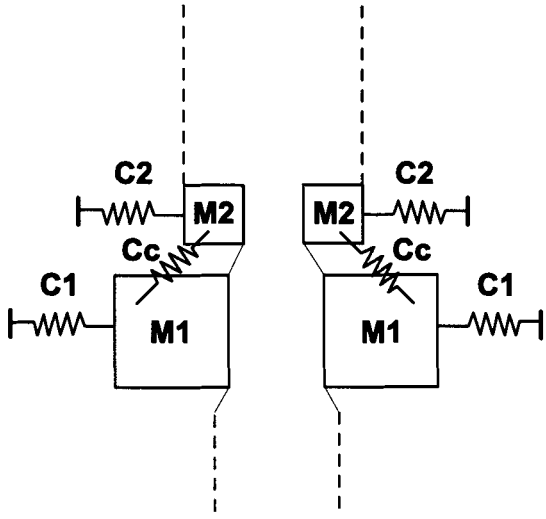

Figura 2.3: *Sistema masa-resorte para el modelado de los tejidos glotales.*

El área mínima entre ambos pliegues vocales determina el área glotal. Esta relación es aproximada, porque la forma anteroposterior de las cuerdas vocales exhibe bastante irregularidad. En este sentido, el área mínima delimitada por las masas M_1 y M_2 de la Figura 2.3 define A_g. El tiempo transcurrido desde que $A_g > 0$ hasta que se alcanza el área máxima se denomina *fase abierta.* El rango de frecuencia natural de este sistema mecánico, para locutores adultos, normalmente oscila entre 100 y 300 Hz. Por consiguiente, para introducir energía en una banda de frecuencia más amplia, el sistema glotal apela al cierre o reducción relativamente abrupta del flujo de aire. El intervalo en el cual ocurre dicha reducción es la *fase de retorno.* Finalmente, durante la *fase cerrada* se verifica $A_g = 0$.

La Ecuación 2.2 describe la dinámica del tejido inferior [78]. Allí, x_1 denota el desplazamiento lateral del tejido, y x_0 es la separación lateral inicial, en reposo, entre las dos porciones inferiores.

$P_s d_1$ es la fuerza promedio por unidad de longitud ejercida sobre el tejido, y el factor d_1 identifica a la longitud vertical promedio de las masas.

$$M_1 \ddot{x}_1 + \frac{1}{C_1}(x_1 - x_0) = P_s d_1 \tag{2.2}$$

Estableciendo $w_0 = 1/\sqrt{M_1 C_1}$, la solución de la Ecuación 2.2 es

$$x_1(t) = (P_s d_1 C_1 + x_0)(1 - \cos w_0 t) \tag{2.3}$$

Nótese que la frecuencia w_0 depende de la masa y del coeficiente de elasticidad. Valores más pequeños de estos parámetros corresponden a un tejido menos voluminoso, y por consiguiente, a una frecuencia vibratoria superior. Para los fines del análisis, la energía supraglotal no incide directamente sobre el tejido superior. Entonces, asumiendo también que $C_2 \gg C_c$, el movimiento del tejido superior se ajusta a la Ecuación 2.4:

$$x_2(t) = x_{20}(1 - \cos w_1 t) \tag{2.4}$$

donde $w_1 = 1/\sqrt{M_2 C_c}$, y x_{20} es el desplazamiento lateral del tejido inferior una vez que alcanza la amplitud x_{10}. En resumen, las constantes determinan la masa y grado de interconexión de los tejidos, y la fuerza ejercida para movilizarlos. Como las señales objeto a invertir provienen de locutores adultos masculinos y femeninos, en el Cuadro 2.1 se reúnen los valores típicos para reproducir en el modelo de dos masas las frecuencias fundamentales pertinentes.

Cuadro 2.1: *Parámetros típicos del modelo glotal para el locutor masculino y el femenino [78].*

PARÁMETRO		MASCULINO	FEMENINO
P_s	(dinas/cm^2)	8000	8000
x_0	(cm)	0.010	0.005
d_1	(cm)	0.200	0.133
M_1	(gm/cm)	0.1	0.04
C_1	(cm^2/dina)	3×10^{-5}	1.9×10^{-5}
M_2	(gm/cm)	0.02	0.008
C_c	(cm^2/dina)	5×10^{-5}	3.1×10^{-5}
x_{10}	(mm)	0.6	0.3
x_{20}	(mm)	0.3	0.015
Área Glotal Máxima (cm^2)		0.107	0.034

De acuerdo con los datos del locutor masculino y con la Ecuación 2.3, la frecuencia natural del tejido inferior es $f_0 = w_0/(2\pi) = 92$ Hz, y $x_1(t)$ alcanza un máximo de 1.2 mm. Sin embargo, por el efecto Bernoulli, este pico teórico en la práctica nunca se alcanza. En el instante en que inicia la separación de los tejidos superiores, la fuerza $P_s d_1$ desciende a cero. Esta caída acontece, aproximadamente, cuando $x_1(t)$ alcanza la amplitud x_{10}, que en el caso masculino vale 0.6 mm. A partir de este punto, $x_1(t)$ sigue un movimiento sinusoidal sin aplicación de fuerza externa, hasta que las masas entran nuevamente en contacto. Como muestra la Figura 2.4, el pico real alcanzado se ubica en torno a los 0.9 mm. Obviamente, cuando se satisface $x_1(t) = x_{10}$, los pliegues superiores inician

su separación (tiempo **B** en la figura), principalmente por efecto del acoplamiento C_c con el tejido inferior. En la Figura 2.4, el movimiento de la masa superior se grafica respetando íntegramente la Ecuación 2.4. Sin embargo, realmente la caída de la curva $x_2(t)$ después del tiempo **C** en que se intersecta con $x_1(t)$ es más abrupta, y no incide en la forma del área glotal. Empero, el comportamiento de $x_2(t)$ hasta el cierre del tejido superior contribuye con la duración de la fase cerrada del pulso glotal. El contacto de las masas superiores, que en la figura acontece a los 9 ms, en la práctica se produce a los 8 ms (tiempo **E**).

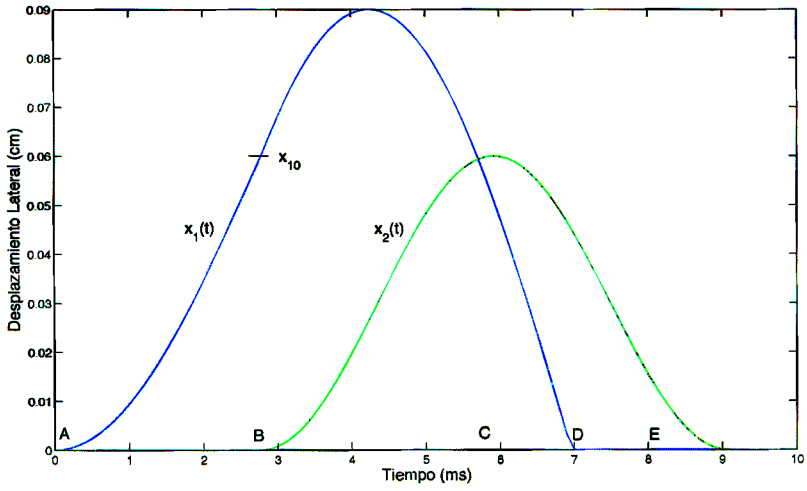

Figura 2.4: *Desplazamiento lateral de las masas inferiores y superiores (locutor masculino).*

$x_1(t)$ y $x_2(t)$ revelan claras diferencias de amplitud y fase, importantes para la definición de las etapas del pulso glotal. Recuérdese que el área glotal es proporcional a la mínima separación entre tejidos de la glotis. Por ende, $A_g = 0$ entre **A** y **B**. Luego de este tiempo, la glotis permanece abierta hasta que los tejidos inferiores se cierran en **D**. La fase de retorno está delimitada por dicho cierre y la intersección de $x_1(t)$ y $x_2(t)$ (tiempo **C**). La glotis permanece cerrada entre **D** y **E**. En la Figura 2.5 se ilustran las tres fases, junto con la curva de área glotal. La fase cerrada tiene una duración de 4 ms, igual a la duración del retorno y la fase abierta combinadas. En consecuencia, la frecuencia fundamental $F0$ de la vibración glotal es 125 Hz, que supera ligeramente a la frecuencia natural f_0 de los tejidos inferiores.

El análisis aplica de forma similar en el caso de la glotis femenina, y las curvas pertinentes se muestran en las Figuras 2.6 y 2.7. En ese caso, $F0 = 250$ Hz.

Para la síntesis articulatoria, la señal de excitación no es el área glotal A_g, sino la velocidad del volumen de aire, U_g [16]. Ambas magnitudes resultan proporcionales, según la Ecuación 2.1. De este modo, la Figura 2.8 presenta la velocidad del volumen para las áreas calculadas.

El análisis evidencia una debilidad fundamental para la síntesis articulatoria. Típicamente, la excitación requerida para articular fonemas sonoros consta de una secuencia de pulsos glotales cuya forma varía entre instancias. Las variaciones, concretamente de amplitud y frecuencia, reciben los nombres de *shimmer* y *jitter*, respectivamente, y contribuyen en gran medida con la *naturalidad* percibida en las emisiones de los sintetizadores articulatorios. Estas variaciones resultan de los procesos no lineales y caóticos en el aparato fonador y del alto nivel de interacción entre sus componentes [23, 40]. El modelado ingenuo de dichas características mediante la concatenación de pulsos glotales modificados aleatoriamente introduce perturbaciones en la señal sintética, con escaso

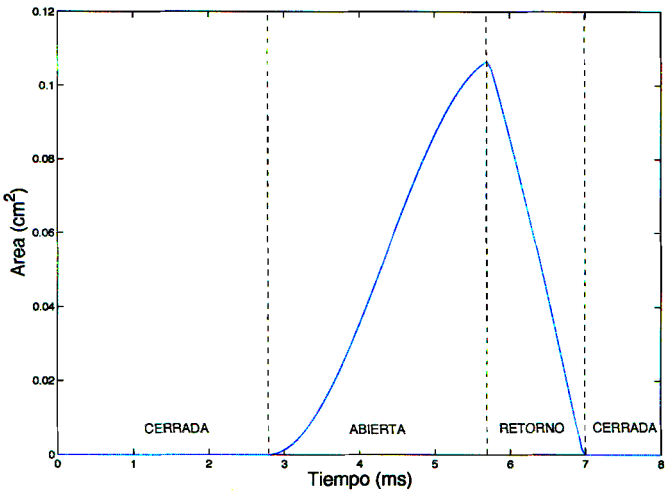

Figura 2.5: *Área glotal (locutor masculino).*

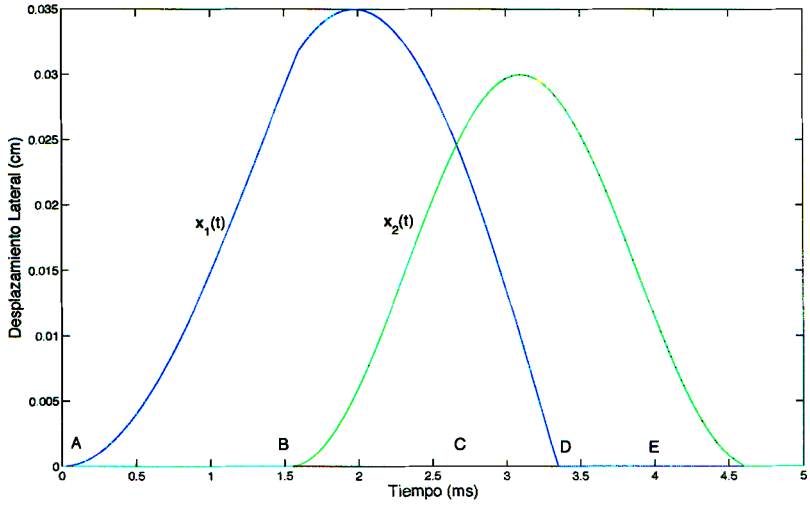

Figura 2.6: *Desplazamiento lateral de las masas inferiores y superiores (locutor femenino).*

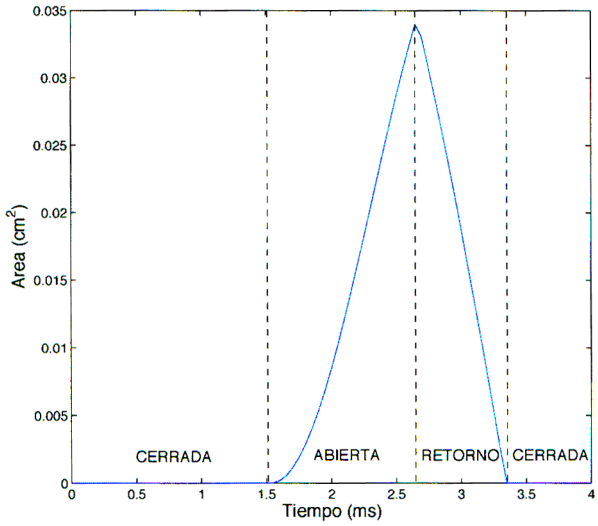

Figura 2.7: *Área glotal (locutor femenino).*

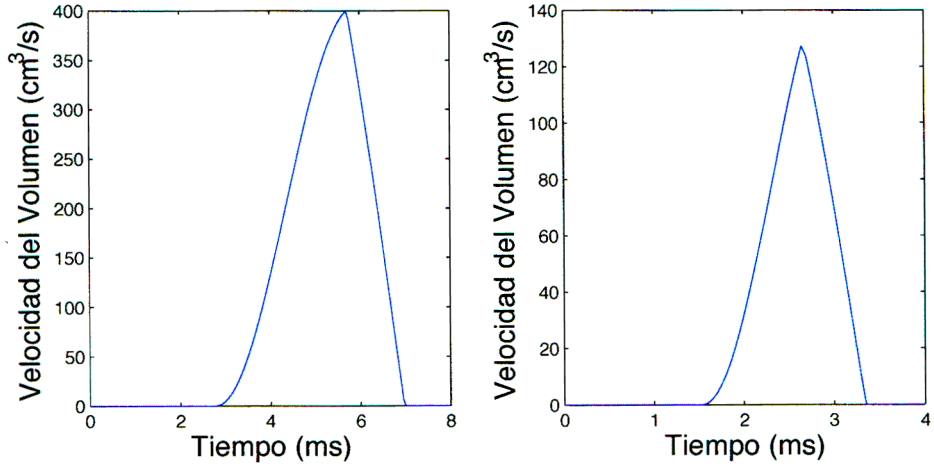

Figura 2.8: *Velocidad del volumen para el locutor masculino (izquierda) y femenino (derecha).*

o nulo provecho en la calidad de la señal [16]. De mayor valor científico resulta utilizar alguna técnica que pueda recuperar la dinámica subyacente en las secuencias de pulsos glotales, y recrearla. En este sentido, las Redes con Estados de Eco han sido empleadas exitosamente en el aprendizaje de dinámicas caóticas y en el modelado de señales cuasi-periódicas [3, 29, 30].

2.3. Redes con Estados de Eco

Las Redes con Estados de Eco (ESN) constituyen un tipo de red neuronal recurrente [30]. Estas redes exhiben por lo menos un ciclo de conexiones sinápticas, y según su teoría matemática, implementan sistemas dinámicos con precisión arbitraria [25]. Aunque dotadas de gran riqueza expresiva, el entrenamiento de estas redes resulta mucho más complejo que el de las variantes no recurrentes. En general, las redes neuronales recurrentes comparten la misma estructura: constan de *neuronas* o *unidades de procesamiento* conectadas por *sinapsis*, y la intensidad de cada vínculo depende de un *peso*. Típicamente la ESN distingue entre unidades de entrada, internas y de salida, presentadas en la Figura 2.9, con niveles de activación respectivos $u_i(n)$, $x_j(n)$ y $y_k(n)$, donde i, j, y k indexan las unidades, y $n = 1, 2, 3, ...$ La dependencia de las funciones de activación en el índice entero n refleja el interés de esta investigación sólo en las redes de tiempo discreto, apropiadas para el procesamiento de señales digitales. Formalmente, una red recurrente incorpora K unidades de entrada, N unidades internas, y L unidades de salida, y por ende los vectores de activación son

$$\begin{aligned} \mathbf{u}(n) &= [u_1(n)\ u_2(n)\ ...\ u_K(n)]^T \\ \mathbf{x}(n) &= [x_1(n)\ x_2(n)\ ...\ x_N(n)]^T \\ \mathbf{y}(n) &= [y_1(n)\ y_2(n)\ ...\ y_L(n)]^T \end{aligned} \tag{2.5}$$

Los pesos asociados a las conexiones entre neuronas se agrupan en las matrices W_u, W_x y W_y, con dimensiones $N \times K$, $N \times N$ y $L \times (K+N+L)$, respectivamente. W_u contiene las conexiones que parten de las K unidades de entrada y arriban a las N unidades internas. Por su parte, W_x almacena las posibles $N \times N$ conexiones entre las unidades internas. A su vez, W_y agrupa las conexiones que parten de todas las unidades de la red y arriban a las unidades de salida. Existe otra matriz, W_b, que aloja las conexiones proyectadas desde las unidades de salida hacia las unidades internas. Así, W_b, de dimensión $N \times L$, actúa como multiplicador de las señales de retroalimentación. El cambio de estado en las unidades internas y de salida viene dado por las Ecuaciones 2.6 y 2.7.

$$\mathbf{x}(n+1) = f_x(W_u\mathbf{u}(n+1) + W_x\mathbf{x}(n) + W_b\mathbf{y}(n) + v(n)) \tag{2.6}$$

$$\mathbf{y}(n+1) = f_y(W_yB(n+1)) \tag{2.7}$$

En la Ecuación 2.7, $B(n+1)$ representa la concatenación de los vectores de activación $\mathbf{u}$(n+1), $\mathbf{x}$(n+1), y $\mathbf{y}$(n). Por otro lado, f_x y f_y son las funciones de activación, normalmente sigmoidales, de las unidades internas y externas, respectivamente. Por su parte, $v(n)$ introduce un ruido blanco de leve magnitud, muestreado a partir de una distribución uniforme sobre $[-0.001, 0.001]$. $v(n)$ es un término requerido en la práctica para lograr la estabilidad de la ESN [29], entendiendo estabilidad como la convergencia de las salidas de la red hacia las señales de entrenamiento.

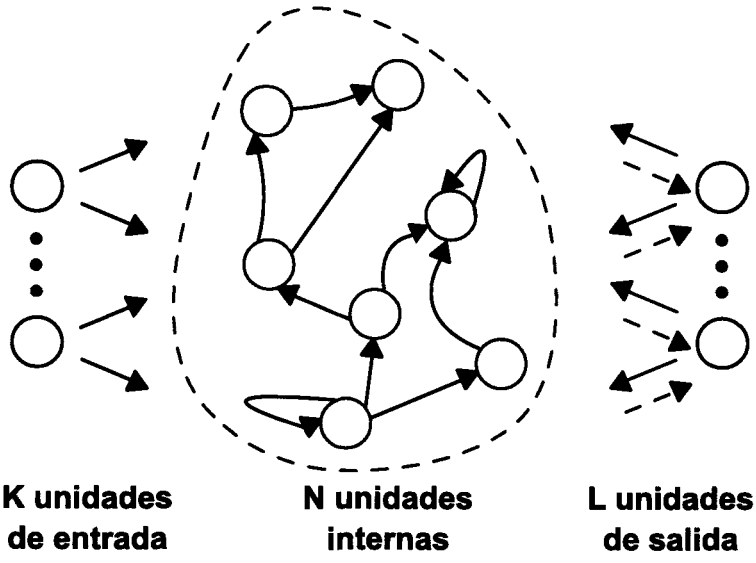

Figura 2.9: *Red con Estados de Eco.*

Un rasgo característico de la ESN es su mecanismo de entrenamiento. En una red de este tipo, sólo los pesos en W_y constituyen incógnitas; los demás pesos se fijan aleatoriamente. De este modo, el aprendizaje consiste en obtener, mediante regresión lineal, los pesos que permitan una buena aproximación de las salidas de la red a las muestras de una señal objeto o de entrenamiento. Esta condición implica que el aprendizaje de las ESN pertenece a las variantes supervisadas. Para los propósitos de este capítulo, la señal de entrenamiento corresponde al pulso glotal U_g revisado anteriormente. El pulso glotal actúa como excitación durante la síntesis articulatoria de fonemas sonoros. En este sentido, la ESN procede como un modelo de caja negra del sistema glotal, aprendiendo su dinámica a partir de la observación de una señal de ejemplo. Después de esta etapa de **Aprendizaje**, sigue la fase de **Explotación**, en la cual la ESN genera una señal de salida según la activación $\mathbf{y}(n)$. Si la red aprendió exitosamente, entonces su salida reproducirá la señal de entrenamiento. Una propiedad de las ESN, que justifica su adopción para el aprendizaje de la señal glotal, es que con los parámetros adecuados la referida reproducción no imita a la perfección la señal de entrenamiento: produce una señal aproximadamente periódica, aún si la señal de entrenamiento posee una periodicidad estricta [3]. De esta manera, la ESN recrearía la perturbación *jitter*.

La distribución de unidades, las funciones de activación y las matrices de pesos constituyen los parámetros principales de la ESN. Una vez definidos, puede desarrollarse la fase de aprendizaje de la red, que comprende dos actividades:

1. **Muestreo**: La red entra en operación, calculando los niveles de activación internos $\mathbf{x}(n+1)$, como indica la Ecuación 2.6. Sin embargo, en cada paso $n+1$ la enésima muestra de la señal de entrenamiento, $U_g(n)$, reemplaza a $\mathbf{y}(n)$. En la literatura de las ESN este proceso suele llamarse *escritura de la señal de entrenamiento en las unidades de salida.*

   El conjunto de unidades internas recibe el nombre de Reservorio Dinámico (DR) porque su matriz de pesos W_x determina la capacidad del sistema para la recuperación de dinámicas. En resumen, el muestreo determina la secuencia de niveles de activación del DR, inducida principalmente por la señal de entrenamiento, por la entrada, y por las matrices W_x y W_b.

2. **Cálculo de Pesos**: El aprendizaje se completa con el cómputo de componentes de W_y para una regresión de las muestras de la señal objeto sobre los estados de activación del DR, lo que equivale a minimizar la diferencia entre las salidas de la red y la señal objeto. Específicamente, se determina el W_y que minimiza el término $(U_g(n+1)-y(n+1))^2$. En las ESN no se establecen conexiones entre las unidades de salida, por lo que puede prescindirse del vector $\mathbf{y}(n)$ en la formación de $B(n+1)$.

   En la práctica, la minimización no considera los primeros n_t niveles de activación obtenidos en el muestreo, porque con esas muestras ($n \leq n_t$) la dinámica de la red se encuentra determinada parcialmente por el estado inicial arbitrario $\mathbf{x}(0)$. n_t se fija empíricamente, de tal manera que garantice que el efecto de $\mathbf{x}(0)$ haya desaparecido en las muestras $n > n_t$, y que sólo se manifiesten los efectos de la señal de entrenamiento escrita en las unidades de salida. Si n_t se establece mal, la red no podrá reproducir la dinámica de la señal objeto.

Obsérvese que durante el muestreo la señal de entrenamiento induce cambios en el estado del DR por retroalimentación a través de W_b. Y luego, con el cálculo de pesos se obtiene la mejor matriz W_y para recrear la señal a partir de sus propios *ecos* $x_j(n)$. De allí proviene el nombre de este tipo de red neuronal.

Un factor determinante para el aprendizaje de la ESN es el **radio espectral** α_r, igual al máximo valor propio de W_x [29]. Si $\alpha_r > 1.0$, la red no tiene estados de eco [30], lo que significa que no puede reproducir la señal de entrenamiento a partir de combinaciones de las unidades del DR. Nótese que la cualidad de estados de eco se define al crearse la red, en la obtención aleatoria de los componentes de W_x, proceso que obviamente depende de las probabilidades y pesos sinápticos. Las investigaciones han demostrado que si α_r es grande (próximo a 1.0) entonces el DR resulta idóneo para el aprendizaje de señales lentas, que cambian a lo largo de muchas muestras, como en el caso de la señal glotal modificada.

En la Explotación se calculan las salidas de la red empleando normalmente las Ecuaciones 2.6 y 2.7. Esta vez, como ya se ha calculado W_y, no hay que reemplazar $\mathbf{y}(n)$ con $U_g(n)$ al calcular $\mathbf{x}(n+1)$.

2.4. Modelado de la Excitación Glotal mediante Redes con Estados de Eco

Una vez revisada la teoría fundamental de las ESN, el resto del capítulo presenta las redes creadas para el aprendizaje de la señal de excitación glotal, incluyendo su definición, entrenamiento y los resultados de la fase de Explotación. La señal de excitación objeto puede construirse por concatenación de pulsos glotales como los mostrados en la Figura 2.8. La Figura 2.10 exhibe una secuencia de 1600 muestras para el locutor masculino. Por cuanto esta señal se utilizará como excitación en la síntesis de fonemas sonoros, su frecuencia de muestreo debe igualarse a la frecuencia de discretización del modelo acústico, que como se indica en el Capítulo 4, asciende a 20 kHz. De las 160 muestras del pulso, la fase cerrada ocupa 76. En esta forma, empero, la señal resulta inabordable para la ESN, por cuanto esta red no puede aprender subsecuencias constantes relativamente largas [29]. Evidentemente, en la Figura 2.10 tales subsecuencias corresponden a la fase cerrada, por lo que ésta se descarta y la señal adquiere la forma de la Figura 2.11. Esta remoción no acarrea graves consecuencias, porque a partir de las otras fases y de F0 puede reconstruirse la fase cerrada antes de proporcionar la señal al sintetizador articulatorio.

Una vez establecida la señal objeto, se prosigue a la definición de parámetros. En primer lugar, se usa la activación sigmoidal típica de las ESN [30]. Y como la señal de excitación es unidimensional, se establece $L = 1$. Además, $K = 0$ porque no se requieren señales de entrada para indicar la

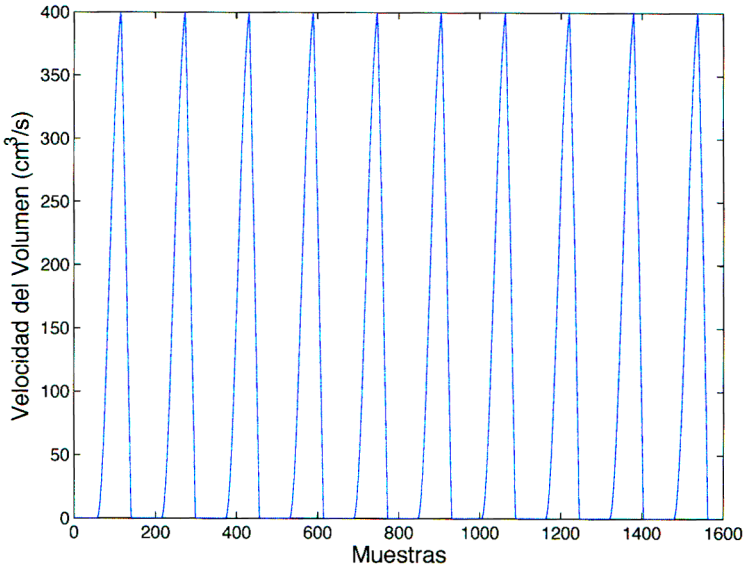

Figura 2.10: *Señal de Excitación Glotal.*

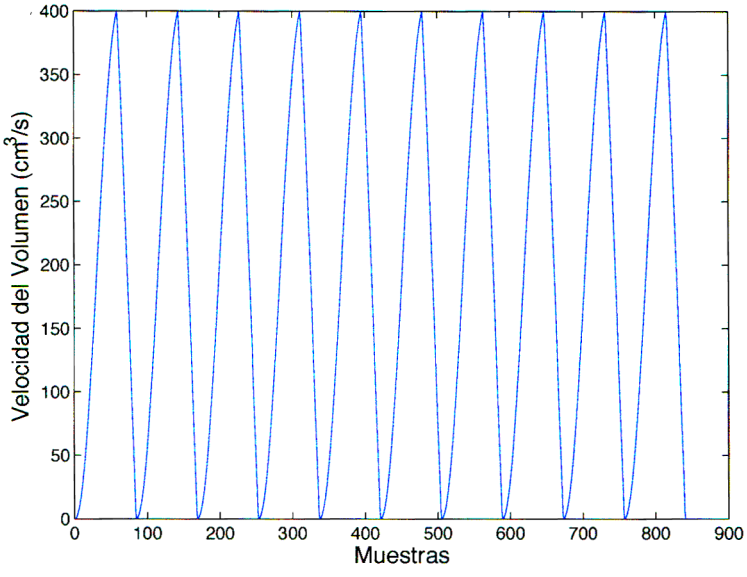

Figura 2.11: *Señal de Excitación Glotal sin Fase Cerrada.*

ocurrencia de algún evento. Por el contrario, la selección de la cantidad de unidades internas plantea el compromiso entre un DR con una pobre capacidad expresiva o el sobreajuste (*overfitting*) de la red si N es muy grande para el tipo de señal de entrenamiento. Normalmente, los problemas complejos exigen una cantidad más alta de unidades internas. En el reporte original de las ESN [29], una red con 400 unidades internas aprendió exitosamente secuencias discretas con período de 48 muestras, más complejas que $U_g(n)$. Entonces, inicialmente se establece $N = 400$, siguiendo los resultados de la referida investigación. Igualmente, la probabilidad de sinapsis entre dos unidades internas se fija en 0.006, con los pesos extraídos aleatoriamente del intervalo $[-0.40, 0.40]$, siguiendo el enfoque presentado en [29]. El próximo paso consiste en ajustar el número de unidades internas, la probabilidad de sinapsis entre dichas unidades, y el rango de pesos sinápticos, utilizando como guía el radio espectral α_r de la red. En este sentido, la probabilidad 0.006 y el intervalo $[-0.40, 0.40]$ garantizan, en promedio, un $\alpha_r = 0.6497$. Aún con este radio espectral, el DR posee capacidad expresiva para intentar la aproximación de $U_g(n)$, si bien la salida exhibe una distorsión severa. En la actualidad, sólo la experimentación manual permite corregir el problema [29, 30]. Para reducir el tiempo de entrenamiento, N se disminuyó a 300, y la probabilidad se incrementó sistemáticamente hasta determinar el valor crítico 0.019. Estos valores condujeron a un α_r que promedió 0.9908, en 20 redes creadas. $N < 300$ ocasionó la pérdida de capacidad expresiva de la red.

Por otra parte, como $\mathbf{y}(n)$ depende de las salidas previas, se requiere retroalimentación sustancial. Así, la probabilidad de sinapsis entre la unidad de salida y las unidades del DR se fija en 0.35, y los pesos de W_b se extraen aleatoriamente del intervalo $[-2.0, 2.0]$ [30]. En resumen, el Cuadro 2.2 agrupa los parámetros de la ESN. Allí, los pasos de escritura representan la cantidad de muestras de la señal objeto escritas en la unidad de salida durante el muestreo. Al igual que con n_t, no existen aún métodos analíticos para su determinación. Sin embargo, como la señal de entrenamiento resulta bastante regular, un número grande de muestras no aporta información nueva al aprendizaje completado en la etapa de cálculo de pesos. En consecuencia, bastan 500 muestras, que abarcan aproximadamente 6 pulsos (ver Figura 2.11). Por su parte, $n_t = 100$ proporciona un tiempo razonable para la desaparición de los efectos de $\mathbf{x}(0)$, lo que en definitiva implica que la minimización en el cálculo de pesos opera sobre 400 muestras. Por último, en teoría $\mathbf{x}(0)$ puede definirse arbitrariamente [30]: en todos los experimentos subsiguientes se ha igualado al vector nulo.

Cuadro 2.2: *ESN para el modelado de la excitación glotal (amplitud constante).*

PARÁMETRO	VALOR
Unidades de Entrada	0
Unidades Internas	300
Unidades de Salida	1
Función de Transferencia	Sigmoidal
Probabilidad de sinapsis internas	0.019
Rango de pesos internos	$[-0.40, 0.40]$
Probabilidad de sinapsis de retroalimentación	0.35
Rango de pesos de retroalimentación	$[-2.00, 2.00]$
Pasos de escritura	500

La Figura 2.12 muestra la salida (curva roja y continua) de una red entrenada con 500 muestras de la señal en la Figura 2.11. Allí, la señal de excitación prototipo es la curva azul y punteada; esta distribución de colores y estilos se mantiene en las gráficas restantes, relativas a las salidas de las redes. Existen varios detalles dignos de mención. En primer lugar, la salida de la red apenas

supera una amplitud de 0.06, mientras que la señal original de entrenamiento alcanza los 400 cm^3/s. Esto sucede porque la señal de entrenamiento se escaló y desplazó para promediar cero. Las ESN, en su condición original, tienen dificultades para aprender señales de gran amplitud, por lo que se recomienda escalar la secuencia antes del aprendizaje. Posteriormente, para generar las señales sintéticas, deben deshacerse los efectos del escalado. En segundo lugar, como se esperaba, la salida iguala a la perfección la señal de entrenamiento en las primeras 500 muestras. A partir de allí, la salida de la red conserva su estabilidad, en el sentido de inexistencia de oscilaciones arbitrarias, reproduciendo aproximadamente la señal objeto, aunque con variaciones en la frecuencia, como se pretendía. Una gráfica estroboscópica como la Figura 2.13 muestra superpuestos los pulsos generados por la red, después de las 500 muestras, evidenciando con mayor claridad las variaciones de frecuencia.

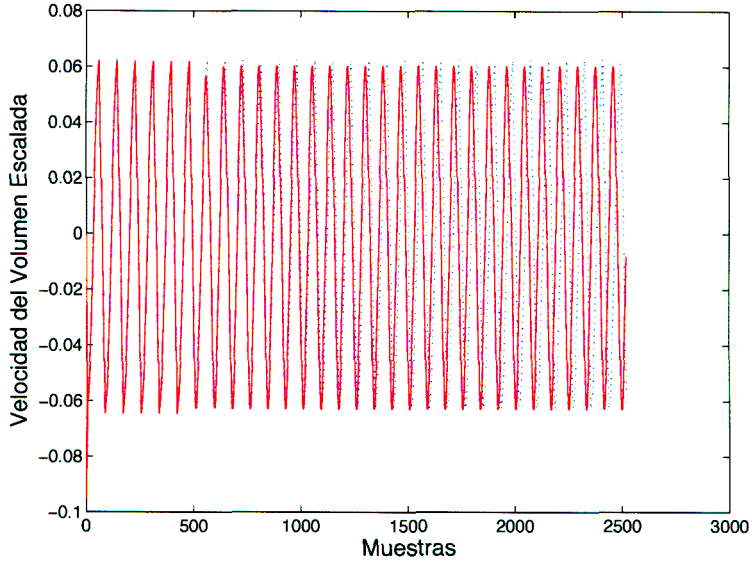

Figura 2.12: *Señal de aprendizaje y salida de la ESN (locutor masculino).*

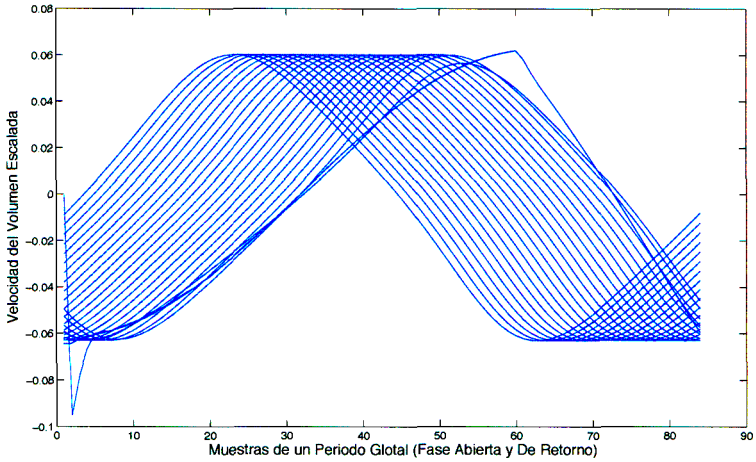

Figura 2.13: *Gráfica Estroboscópica de los Pulsos Glotales generados por la ESN (locutor masculino).*

Una red con idéntica configuración pudo adaptarse sin problemas a la señal construida a partir del pulso glotal femenino en la Figura 2.8. Las gráficas de aprendizaje y de superposición de pulsos se exhiben en las Figuras 2.14 y 2.15, respectivamente.

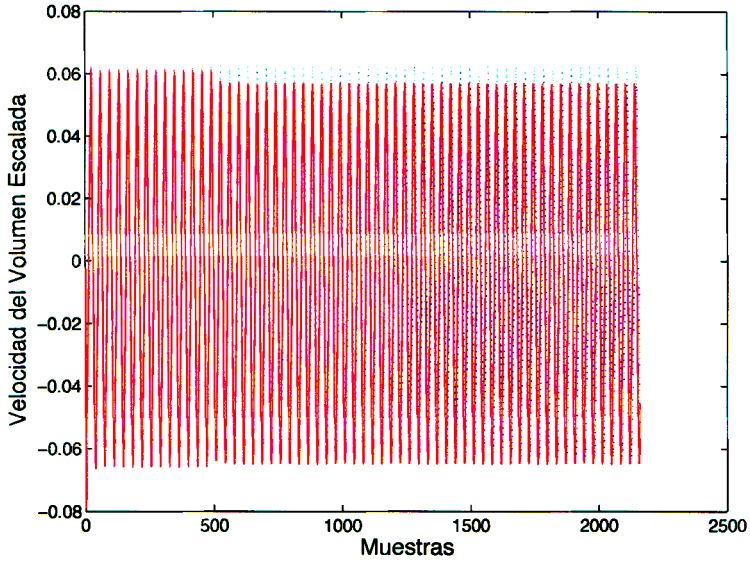

Figura 2.14: *Señal de aprendizaje y salida de la ESN (locutor femenino).*

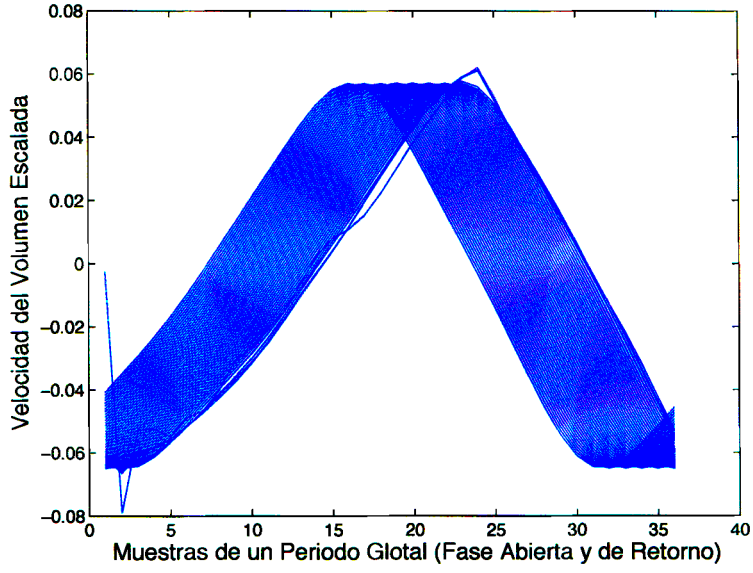

Figura 2.15: *Gráfica Estroboscópica de los Pulsos Glotales generados por la ESN (locutor femenino).*

Como se observa, la red introdujo exitosamente la perturbación *jitter* con las señales de ambos locutores prototipo. Sin embargo, el shimmer no ha logrado recrearse. No se aprecian diferencias de amplitud entre los pulsos generados por la ESN[1]. Para solventar esto, se agrega a la red una unidad

[1]Las diferencias de amplitud respecto a la señal original no resultan de importancia, por cuanto pueden compensarse

de entrada que actúa como receptora de una señal controladora del nivel de amplitud de la salida. La idea es que cuando esta señal controladora cambie, la ESN responda modificando apropiadamente la amplitud de su salida. Ciertamente, en este caso la variación se induce explícitamente mediante un estímulo externo (a diferencia del *jitter* implícito), pero al menos permite simular las variaciones de amplitud, imprescindibles para la naturalidad de la señal sintética. La señal de entrada puede restringirse a cualquier intervalo arbitrario. No obstante, investigaciones precedentes recomiendan una entrada en torno a 3.0, con la finalidad de que el DR trabaje en un rango altamente no lineal de sus unidades sigmoidales [29]. En este sentido, se define como rango de entrada, para el entrenamiento, el intervalo $[3.0, 3.25]$. Se asignará el valor 3.0 a la máxima amplitud, y 3.25 a la mínima.

Ante la mayor complejidad del problema, la cantidad de unidades internas y los pasos de escritura deben incrementarse. Para mantener α_r próximo a 1.0, la probabilidad de sinapsis internas se reduce a 0.011. Las probabilidades y el intervalo de pesos para las sinapsis de entrada se establecen en 0.005 y $[-0.30, 0.30]$, para que el DR reciba la información del exterior [29]. El resto de parámetros de la red se conservan iguales; el Cuadro 2.3 los recopila en su totalidad.

Cuadro 2.3: *ESN para el modelado de la excitación glotal (amplitud variable).*

PARÁMETRO	VALOR
Unidades de Entrada	1
Unidades Internas	500
Unidades de Salida	1
Función de Transferencia	Sigmoidal
Probabilidad de sinapsis de entrada	0.005
Rango de pesos de entrada	[-0.30, 0.30]
Probabilidad de sinapsis internas	0.011
Rango de pesos internos	[-0.40, 0.40]
Probabilidad de sinapsis de retroalimentación	0.35
Rango de pesos de retroalimentación	[-2.00, 2.00]
Pasos de escritura	5000

La Figura 2.16 muestra la señal de entrada, mientras que la Figura 2.17 presenta las señales de entrenamiento y salida de la ESN, para el locutor masculino. Nuevamente la red logra imitar la señal de entrenamiento en las muestras de escritura. En las otras muestras, como la señal de entrada decrece, la ESN responde incrementando la amplitud. Nótese que aún cuando la señal controladora abandona el rango empleado durante el aprendizaje, la red aproxima la señal de entrenamiento con propiedad. Además, redes con la misma configuración del Cuadro 2.3 lograron aprender las señales objeto con descenso de amplitud para el locutor prototipo masculino (Figura 2.18), y de ascenso y descenso para el locutor femenino (Figuras 2.19 y 2.20), utilizando las señales de entrada y pasos de escritura pertinentes.

En conclusión, la ESN logra introducir las variaciones de frecuencia entre pulsos, en todos los ensayos y de forma implícita. Adicionalmente, se logra aprender la asociación entre la señal de entrada y la dinámica del DR, reflejada en la salida de la red. De este modo, también puede inducirse la variación en la amplitud de los pulsos a través del estímulo externo, contribuyendo aún más con la naturalidad de las emisiones sintéticas.

Finalmente, recuérdese que la salida de las ESN se empleará como fuente de energía para la producción de fonemas sonoros, proceso abordado posteriormente en la sección 5.5. De acuerdo con

durante la reconstrucción de la excitación del sistema articulatorio.

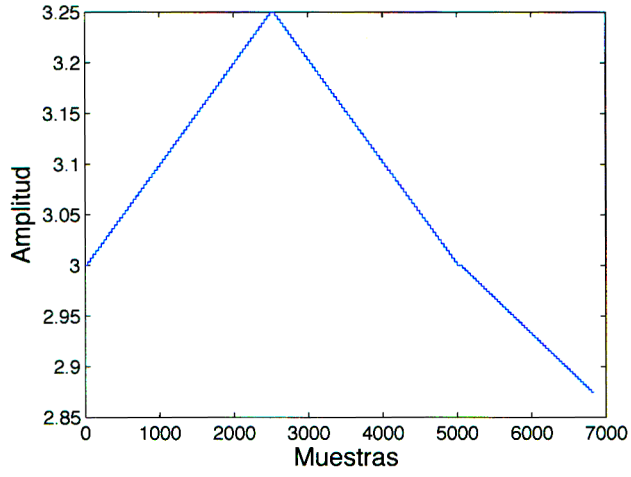

Figura 2.16: *Señal de entrada para el control de la amplitud de la ESN.*

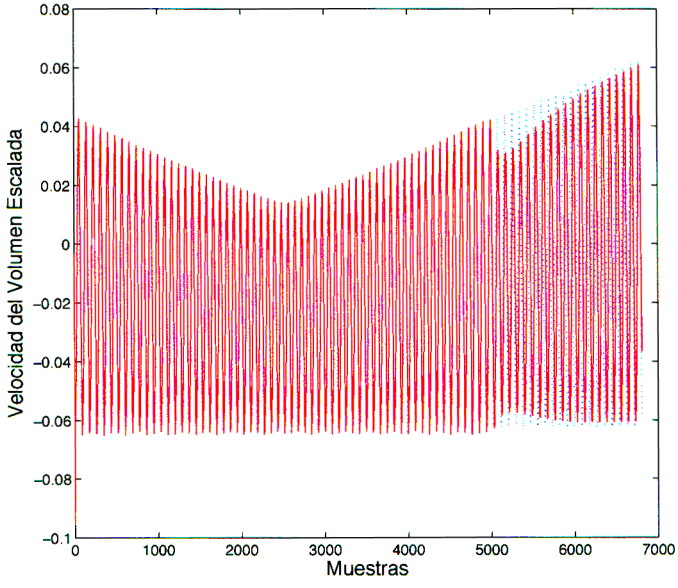

Figura 2.17: *Aprendizaje del ascenso en la amplitud (locutor masculino).*

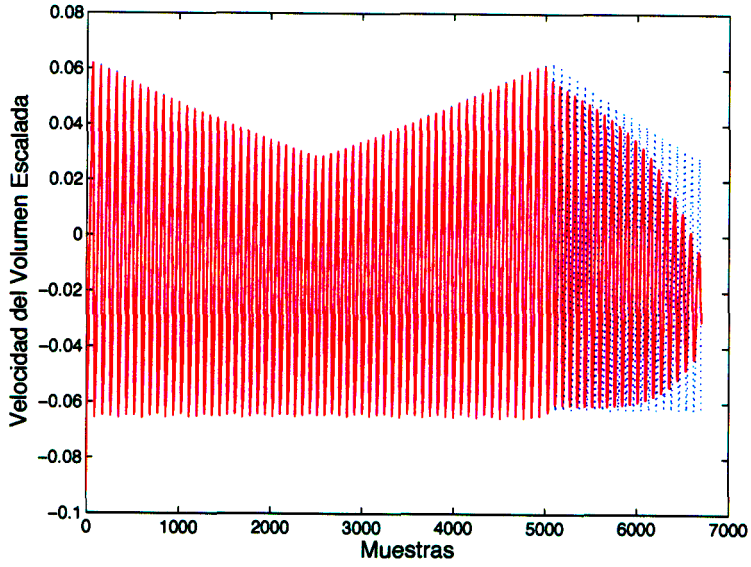

Figura 2.18: *Aprendizaje del descenso en la amplitud (locutor masculino).*

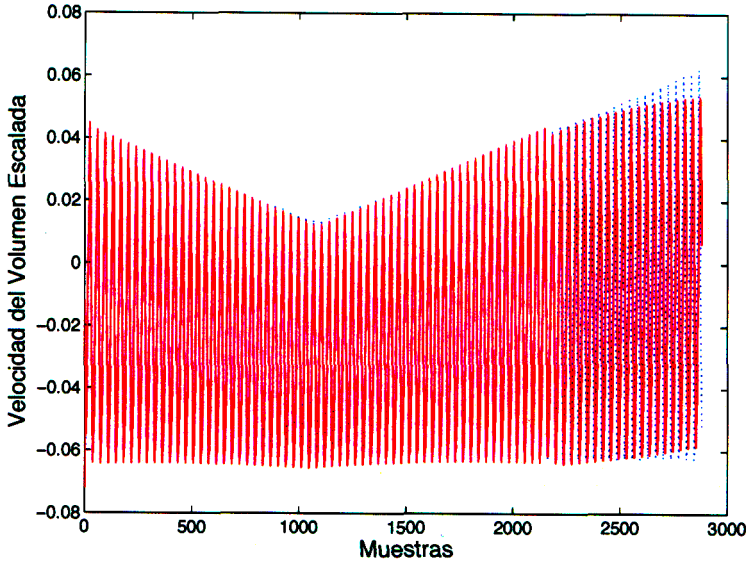

Figura 2.19: *Aprendizaje del ascenso en la amplitud (locutor femenino).*

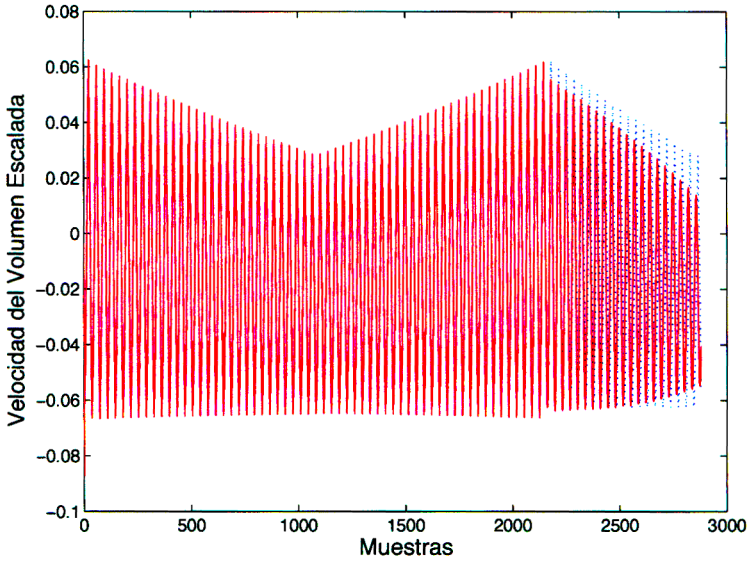

Figura 2.20: *Aprendizaje del descenso en la amplitud (locutor femenino).*

el Modelo Fuente-Filtro discutido en la sección 1.3, esta energía excita los tractos supraglotales para generar la señal de voz. Pero en última instancia, las peculiaridades acústicas de la señal de voz se encuentran determinadas por las configuraciones de dichos tractos durante la articulación. En este sentido, el capítulo siguiente discute el modelo articulatorio construido para representar la geometría supraglotal.