mapeo muchos-a-uno, ya discutido. En este caso, la retracción involucra el músculo MC al máximo. Ciertamente, para las vocales altas el MC puede combinarse con otros músculos no considerados en el modelo articulatorio, como el estilohioideo, para elevar la lengua en las vocales altas [83]. No obstante, el comportamiento del MC en el modelo equivale aproximadamente a su función natural: estrechar el tracto faríngeo durante la deglución. Esto indica que el AGC ha favorecido una configuración que utiliza el MC para compensación articulatoria, es decir, el algoritmo alcanzó un mínimo local, o el modelo articulatorio no puede representar normalmente las características acústicas de la grabación de F_1. Por otra parte, ambos vectores exhiben actividad espuria y contracciones simultáneas de músculos antagonistas, que sin embargo no resultan de importancia porque seguramente disminuirán bajando el umbral de error del AGC. En todo caso, las evaluaciones subjetivas de la sección 5.6 confirman adicionalmente la inversión.

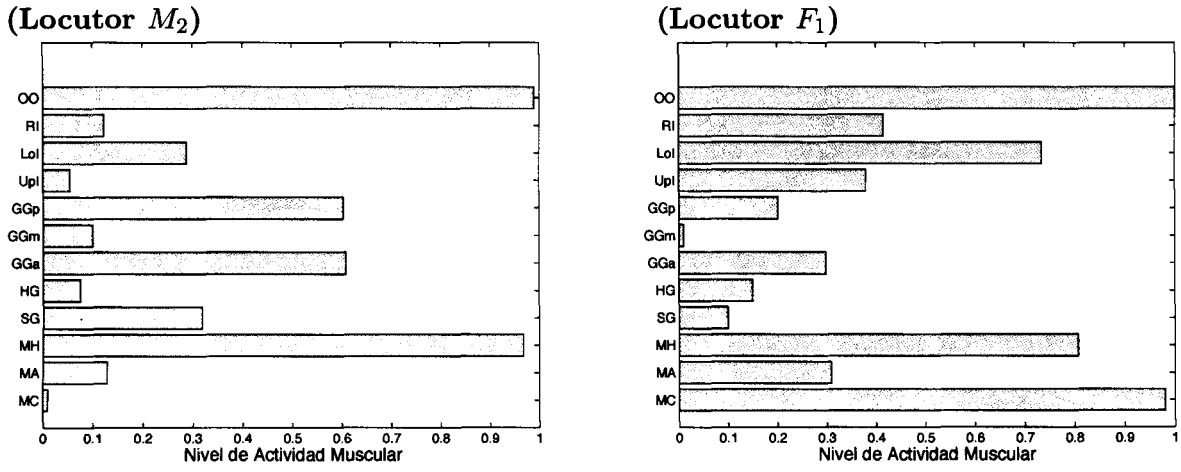


Figura 5.10: *Actividad muscular vinculada a las configuraciones articulatorias recuperadas (vocal /u/)*

5.5.3. INVERSIÓN DEL RESTO DE VOCALES

Con el resto de las vocales la inversión procede similarmente al caso de la /u/, con la función objetivo inalterada. Simplemente, se ejecuta el AGC 20 veces con cada una de las otras vocales del corpus. La Figura 5.12 contiene la evolución del promedio y mejor valor de la función objetivo para las señales invertidas con el menor error, mostrando sólo la mejor inversión por vocal. Por ejemplo, la mejor configuración para la vocal /e/ se obtuvo durante la inversión de la grabación de esa vocal por parte de M_2. Así, sólo se grafican los datos de la inversión de la /e/ grabada por dicho locutor. La Figura 5.12 comprueba la escogencia acertada de los parámetros del AGC: todos los ensayos convergen, y en un número muy bajo de generaciones. Las configuraciones articulatorias respectivas se agrupan en la Figura 5.13. Sin embargo, el Cuadro 5.1 reúne las frecuencias formantes asociadas a la mejor configuración de todas las vocales y locutores.

En una primera aproximación, estas configuraciones para las vocales exhiben consistencia articulatoria. Nótese que en la /a/ el cuerpo de la lengua alcanza una posición relativamente central, y la configuración, en general, corresponde a un tracto vocal abierto. Por el contrario, la /e/ y la /i/ establecen una constricción superior y anterior, en los grados respectivos. Por último, la /o/ implica una constricción posterior, con los labios formando un tubo acústico anterior. Compárense dichos resultados con las resonancias de la Figura 5.11, que si bien no provienen de un locutor de habla

española, resultan muy próximas a las configuraciones de las vocales invertidas aquí, y cumplen apropiadamente un rol ilustrativo.

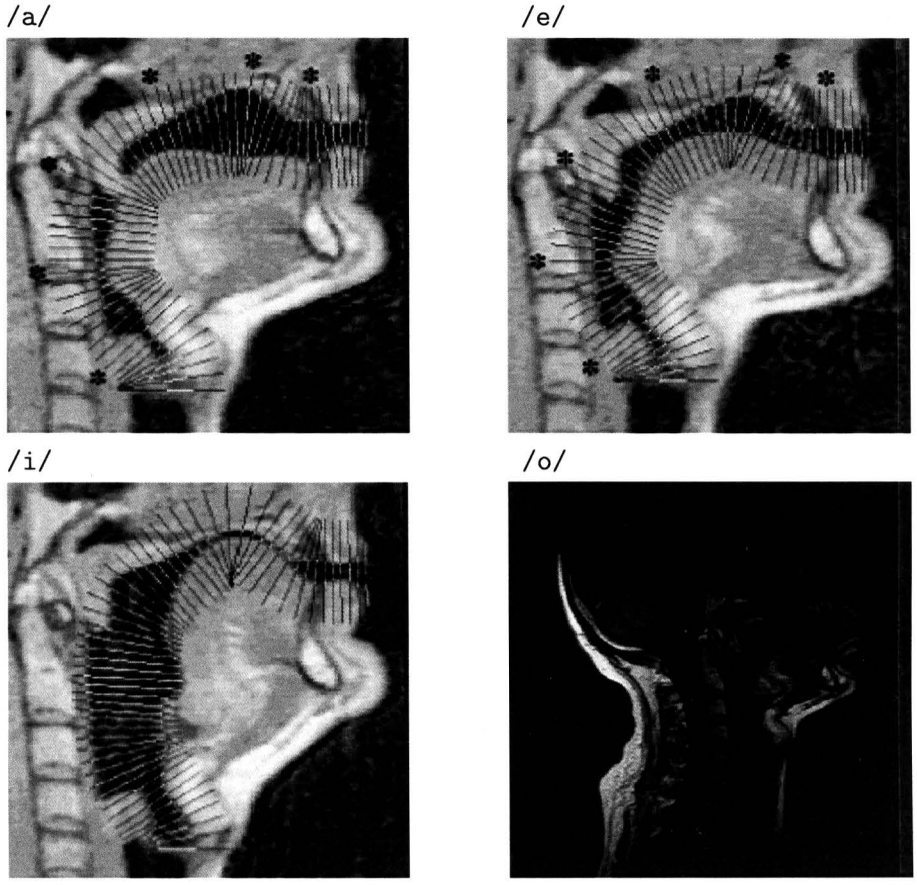
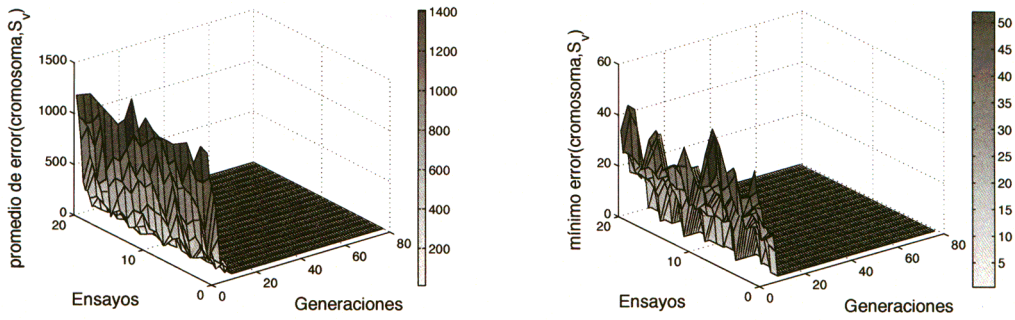


Figura 5.11: *Resonancias magnéticas (otras vocales) [34, 35].*

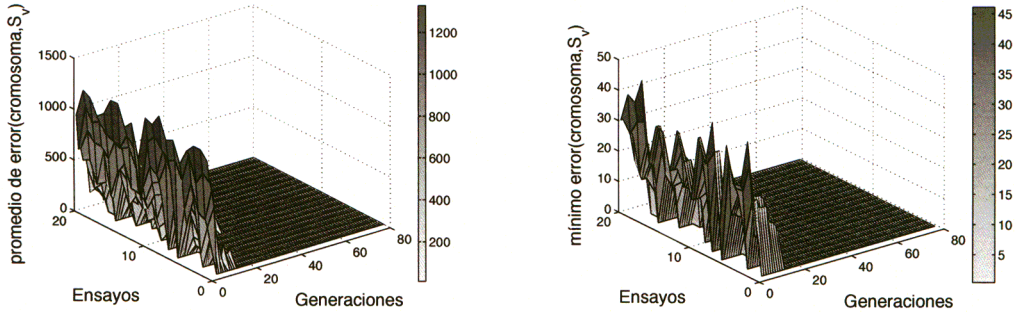
Como indica la Figura 5.14, en todas las vocales actúa MH. Otros niveles satisfactorios son el OO en la /o/, y el RI en la /a/ y la /e/. En la /i/, no obstante, hay demasiada acción del OO, tal vez por compensación. Respecto a la ubicación y forma de la lengua, se tiene:

- /a/: La evolución prefiere un tubo acústico ancho en su parte anterior, como señala la función de área. Para ello, recurre al LoI, GGm y GGa. Y con el MC estrecha el tracto faríngeo.
- /e/: Se acerca la lengua hacia el paladar, combinando la acción ascendente del GGp, SG y MA.
- /i/: Esta vocal ha resultado difícil de formar para el modelo articulatorio, porque cuando el punto $B(x, y)$ asciende, el dorso de la lengua sube en consecuencia. Y como el área se reduce en una zona más anterior que la de la /e/, debe apelar a muchas actividades compensatorias para evitar que el dorso y el ápice colisionen con la frontera superior. Por eso la actuación considerable del GGa, GGm y HG. El GGp y el MA se ocupan del ascenso. Como se trata de una vocal muy anterior, SG es casi nulo.
- /o/: Aquí el HG y el GGa se ocupan de la retracción y depresión de la lengua.

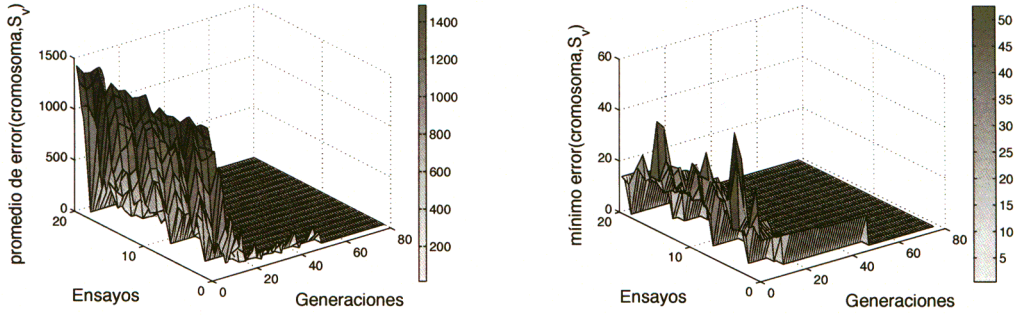
/a/ (Locutor M_1)



/e/ (Locutor M_2)



/i/ (Locutor F_2)



/o/ (Locutor M_1)

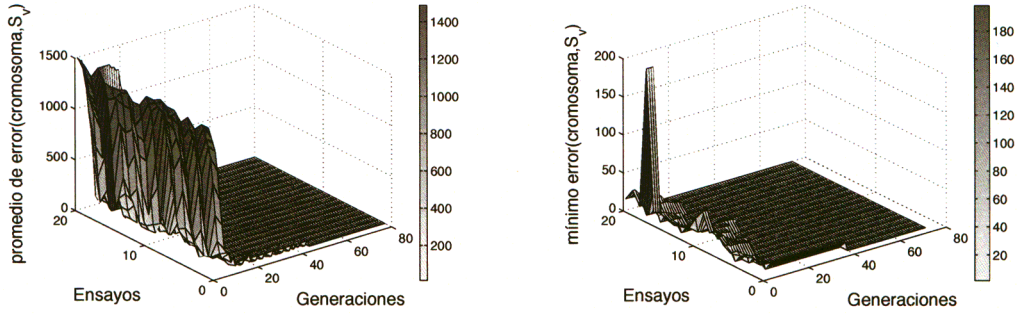


Figura 5.12: *Promedio y mejor valor de la función objetivo por generación (otras vocales).*

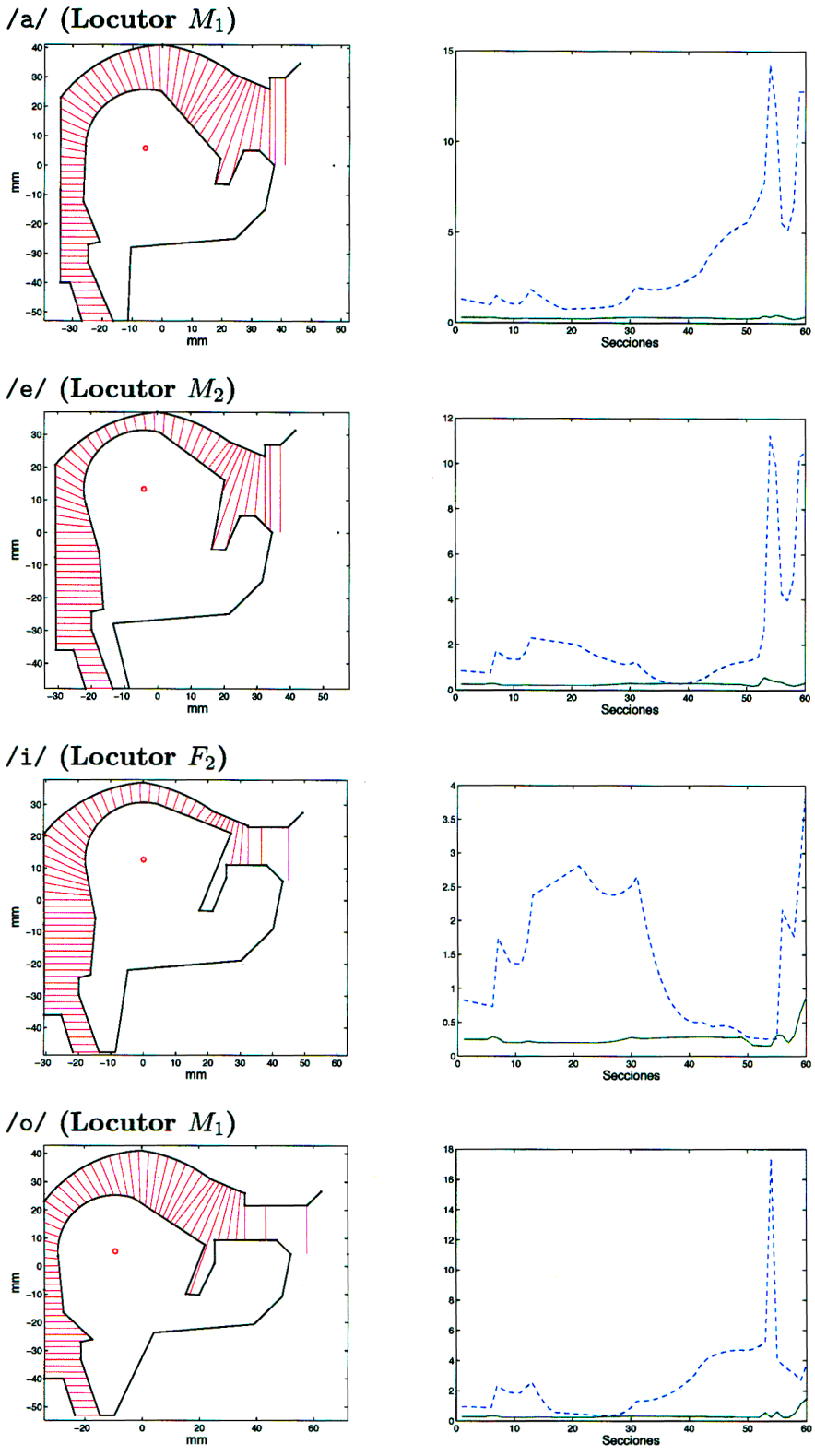


Figura 5.13: *Mejores configuraciones articulatorias recuperadas (otras vocales). A la derecha de cada configuración medial se muestran las funciones respectivas de área (punteada, en cm^2) y longitud (continua, en cm).*

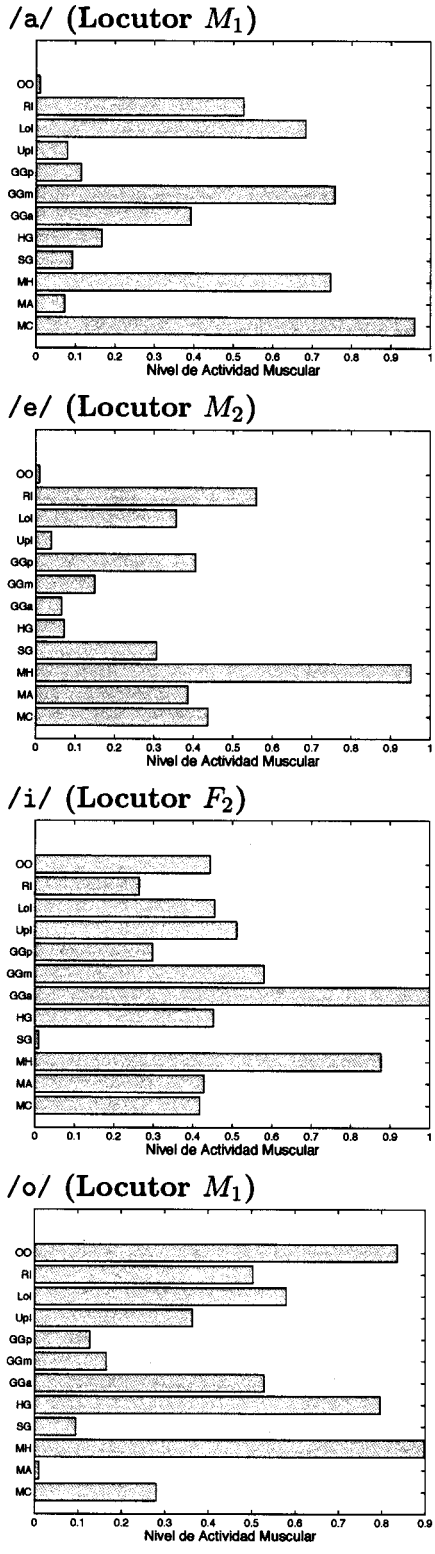


Figura 5.14: *Actividad muscular vinculada a las configuraciones articulatorias recuperadas (otras vocales)*

	Formantes				
Vocal /a/	F1	F2	F3	F4	Error
M_1	0.7988	1.4213	2.4975	3.4856	10.8134
	0.7712	1.4420	2.4428	3.7653	
M_2	0.7925	1.4760	2.3873	3.9031	11.8936
	0.7510	1.4768	2.4655	3.7463	
F_1	0.8874	1.5213	2.8205	4.2674	11.5170
	0.8485	1.5298	2.7057	4.3160	
F_2	0.9445	1.5077	3.3629	4.0228	11.3547
	0.9813	1.4844	3.4342	4.7673	
Vocal /e/					
M_1	0.4683	2.0135	2.6357	3.5555	11.6299
	0.4700	1.8700	2.4715	3.6202	
M_2	0.4725	1.0579	1.8777	2.7157	10.0651
	0.4710	1.0341	1.8702	3.1223	
F_1	0.4663	2.4128	3.0445	3.8602	10.4523
	0.4559	2.4162	2.7424	3.8025	
F_2	0.4563	2.3992	3.0407	3.8715	11.9409
	0.4558	2.4082	3.3483	4.4476	
Vocal /i/					
M_1	0.2841	2.0846	2.9893	3.3611	11.9971
	0.2804	2.0148	2.5842	3.2589	
M_2	0.3553	2.0417	2.6460	3.2197	11.5201
	0.3585	2.0190	2.3590	3.2592	
F_1	0.3646	2.3830	3.2722	3.9065	11.6818
	0.3615	2.4717	2.9292	3.7454	
F_2	0.5041	2.5262	3.2874	4.0300	10.8201
	0.5047	2.4548	3.3451	4.6269	
Vocal /o/					
M_1	0.5408	0.9027	2.5391	3.3097	10.7109
	0.5314	0.8961	2.4065	3.6465	
M_2	0.5388	1.4674	2.2142	3.2293	11.5675
	0.5418	1.4545	2.5053	3.4254	
F_1	0.6038	0.9080	2.4894	3.9899	10.7871
	0.6114	0.9527	2.5985	4.2816	
F_2	0.5709	1.5430	2.9594	3.9718	11.3187
	0.5719	1.5160	2.6163	3.8812	
Vocal /u/					
M_1	0.3847	0.7801	2.4165	3.0608	11.5093
	0.3914	0.7817	2.4604	3.8183	
M_2	0.4306	0.9956	1.7921	3.2270	8.4421
	0.4316	0.9615	1.8016	2.9324	
F_1	0.3826	0.7235	2.4629	3.5756	10.6721
	0.3818	0.7217	2.8024	4.0321	
F_2	0.4407	1.1780	1.8343	2.6685	10.7618
	0.4398	1.1803	1.9206	3.4210	

Cuadro 5.1: *Evaluación de la función objetivo con las mejores configuraciones recuperadas (vocales). Los formantes se expresan en kHz.*

En todos los vectores hay actividades espurias y antagonistas, y en consecuencia, insignificantes. En conclusión, las configuraciones aprendidas resultan consistentes articulatoriamente sobre el plano medial. Además, según el Cuadro 5.1, todas alcanzaron el umbral de error establecido, y los formantes asociados resultan bastante próximos a los formantes de las señales objeto, validando el proceso de inversión desde el punto de vista objetivo, numérico. No obstante, la validación acústica completa de las configuraciones amerita la síntesis de las señales, y la subsiguiente evaluación subjetiva.

5.5.4. INVERSIÓN DE CONSONANTES NASALES

Estas consonantes se producen cerrando completamente el tracto oral en algún punto OC, y abriendo el paso velofaríngeo para admitir flujo de aire en el tracto nasal. La información acústica que identifica a estas consonantes, durante el murmullo nasal, reside en las frecuencias bajas ($<$ 2kHz), mientras que en la transición hacia el próximo sonido se desarrollan algunos eventos acústicos en las frecuencias altas [78]. Como la inversión abordada aquí es de tipo estático, sólo se analiza la región de baja frecuencia. Los picos en dicha región pueden aproximarse con el mismo método basado en PL de la sección 5.3, pero incrementando el ancho de banda admitido para los picos a 1 kHz, con el fin de compensar el efecto de las antiresonancias. Por otra parte, el análisis teórico revela que la primera antiresonancia suele ubicarse alrededor de los 500 Hz, y resulta por efecto del seno maxilar [41,43,71]. La próxima antiresonancia proviene del acoplamiento con el tracto oral. La distancia desde la bifurcación velofaríngea hasta el punto de oclusión es mayor en la /m/ que en la /n/, y por consiguiente, el cero se ubica antes en la /m/. En concreto, el cero reside cerca de 1.2 kHz para la /m/, y de 1.8 kHz para la /n/ [41,78]. Con base en esta información, la función objetivo (o error asociado a un cromosoma) se modifica a

$$error(cromosoma, S_v) = f_1(cromosoma, S_v) + f_2(cromosoma) + W_C^T \left|\alpha_C(\phi(p(t))) - C\right| \quad .$$

Naturalmente, para la /m/, $C = [500\ 1200]^T$, y para la /n/, $C = [500\ 1800]^T$. α_C es una función que retorna los ceros asociados al vector articulatorio $p(t)$, según lo explicado en la sección 5.2. Se calculan sólo tres picos espectrales de igual importancia. De este modo, $W = [10\ 10\ 10]^T$ y $W_C = [10\ 10]^T$. El resto de parámetros del AGC, y el procedimiento en general para la inversión, permanece igual al de las vocales, a excepción de la penalización a las configuraciones con solapamiento de fronteras, que aquí no aplica. Antes de la inversión, el punto E_6 del modelo articulatorio se ubicó a 0.5 cm de E_5, para simular el descenso del paladar blando. En promedio, el AGC converge luego de 31 generaciones. Nuevamente, se presentan los resultados de las señales /m/ y /n/ mejor invertidas, en la Figura 5.15. Los vectores articulatorios respectivos aparecen en la Figura 5.16, y la evaluación de la función objetivo en el Cuadro 5.2.

Como se esperaba, la oclusión de la /m/ es más anterior, en los labios específicamente, y la de la /n/ acontece en la zona alveolar. La superposición de las fronteras superior e inferior en las figuras proviene de la incapacidad del modelo medial para representar la colisión y elasticidad de los tejidos.

La actividad muscular destaca por su claridad. En la /m/, el OO y el MA se combinan para unir los labios, mientras que en la /n/ los músculos intrínsecos elevan la punta de la lengua hacia los alvéolos.

5.5.5. INVERSIÓN DE CONSONANTES FRICATIVAS

El espectro de las fricativas /f/ y /s/ difiere significativamente del de los fonemas estudiados hasta ahora. La turbulencia generada en el tracto oral desvanece la nitidez de las resonancias. Sin embargo, la energía acústica se distribuye siguiendo patrones particulares. En la /f/, la distribución

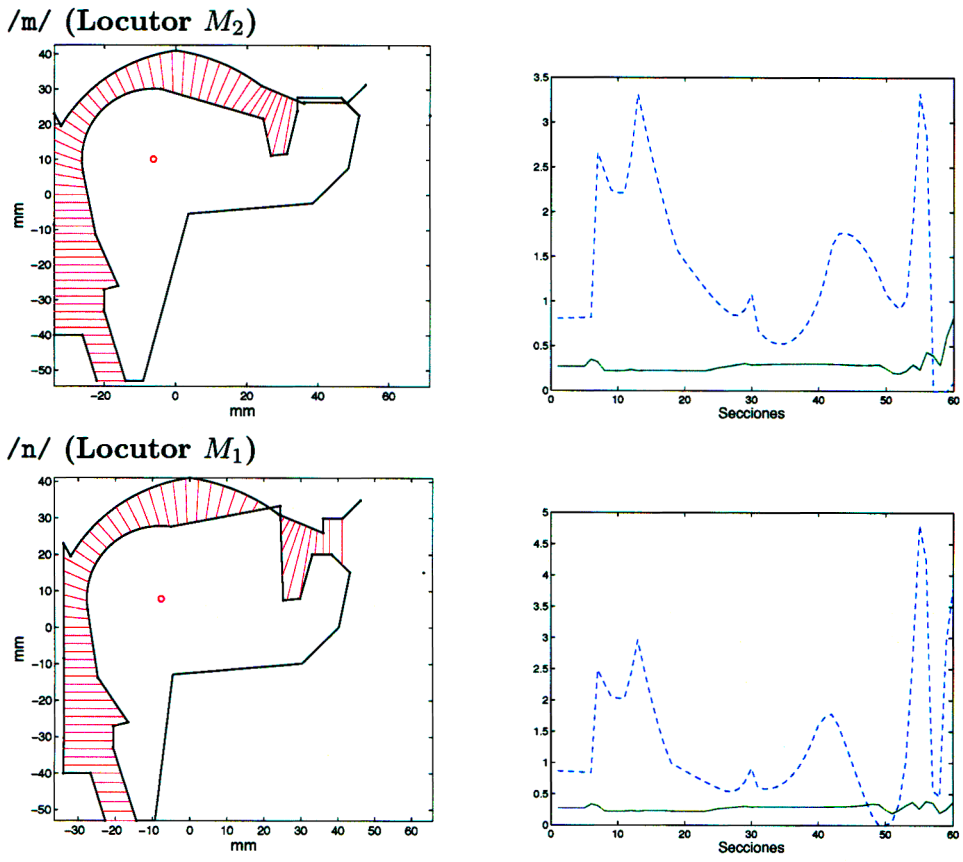


Figura 5.15: *Mejores configuraciones articulatorias recuperadas (consonantes /m/ y /n/). A la derecha de cada configuración medial se muestran las funciones respectivas de área (punteada, en cm^2) y longitud (continua, en cm).*

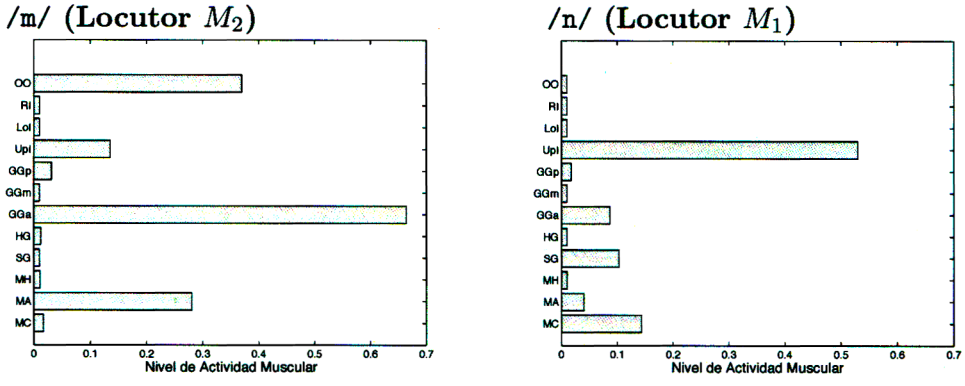


Figura 5.16: *Actividad muscular vinculada a las configuraciones articulatorias recuperadas (consonantes /m/ y /n/)*

	Resonancias			Antiresonancias		
Consonante /m/	P1	P2	P3	C1	C2	Error
M_1	0.2144	0.5940	0.9987	0.5000	1.2000	11.2111
	0.3344	0.5806	1.0555	0.5403	0.9998	
M_2	0.1917	0.6272	0.9520	0.5000	1.2000	11.0355
	0.1944	0.8667	1.5672	0.6034	1.5470	
F_1	0.2018	0.4123	0.9969	0.5000	1.2000	11.4574
	0.3487	0.5712	1.1220	0.5033	1.8332	
F_2	0.3621	0.4099	0.8999	0.5000	1.2000	11.1393
	0.4501	0.6096	1.0133	0.5600	1.1975	
Consonante /n/						
M_1	0.2809	0.6943	0.8473	0.5000	1.8000	10.7975
	0.4002	0.7721	0.9940	0.5493	1.7834	
M_2	0.1515	0.4184	0.8016	0.5000	1.8000	11.8670
	0.2566	0.5262	0.9918	0.6007	1.9088	
F_1	0.2510	0.5007	0.9716	0.5000	1.8000	10.8832
	0.3266	0.5174	0.8096	0.5455	1.8905	
F_2	0.2989	0.5415	0.9960	0.5000	1.8000	10.9954
	0.2866	0.6566	1.1237	0.5078	1.7697	

Cuadro 5.2: *Evaluación de la función objetivo con las mejores configuraciones recuperadas de la /m/ y la /n/.*

resulta relativamente uniforme, mientras que la /s/ evidencia un ascenso alrededor de los 5kHz [78, 89]. La densidad de energía espectral [58] de dos señales del corpus revela este comportamiento (Figura 5.17). Las señales se suministran a un filtro paso alto con 1 kHz como frecuencia de corte [72].

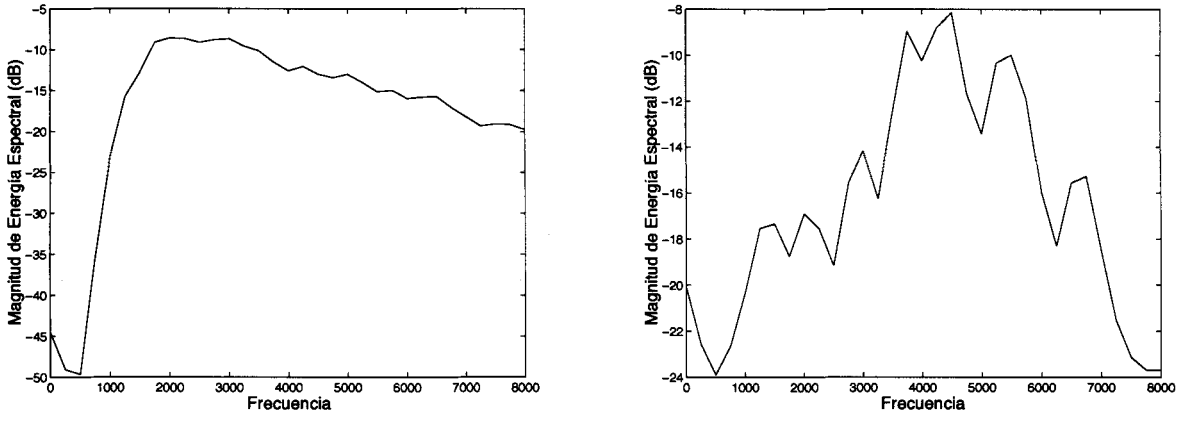


Figura 5.17: *Densidad de Energía Espectral para fricativas. A la izquierda, la /f/. En la derecha, la /s/.*

Una vez más, recuérdese la discusión en la sección 1.4. Un tubo acústico más corto implica unas frecuencias de resonancia más altas. Ocurre que el tubo acústico al frente de la constricción en la /s/ exhibe una resonancia alrededor de los 5 kHz, y por tal motivo la energía espectral se concentra en esa región. Por su parte, la longitud efectiva de la cavidad acústica al frente de la constricción de la /f/ es menor: suponiendo dicha longitud igual a 0.9 cm, la energía se ubica en torno a los 10 kHz [78]. No obstante, los puntos E_{10} y E_{11} del modelo articulatorio se encuentran separados 0.5

cm en la configuración de equilibrio, por lo cual la primera resonancia se ubica alrededor de los 18 kHz. Ante la falta de información en esas frecuencias, la inversión de la /f/ no empleará información derivada del análisis de la señal objeto, sino que se intenta aproximar directamente la concentración de energía en 18 kHz. Por ende, no se utilizarán las señales /f/ del corpus en la inversión. En este caso, se efectúan dos experimentos de inversión, utilizando las configuraciones articulatorias masculina y femenina.

Claramente, la función objetivo debe modificarse. Se trabajará con una sola resonancia, la del tubo acústico al frente de la zona de constricción, por lo que α arroja dicha resonancia. La forma general del error permanece como en la Ecuación 5.8. Sin embargo, como se calcula una sola resonancia, $W = [200]$.

Las configuraciones buscadas deben tener una constricción, no una oclusión, por lo que se penalizan las configuraciones con solapamiento de fronteras. El resto del AGC permanece idéntico. Los resultados están en las Figuras 5.18 y 5.19, y en el Cuadro 5.3. Todos los ensayos alcanzaron el umbral de error, en promedio, después de 16 generaciones.

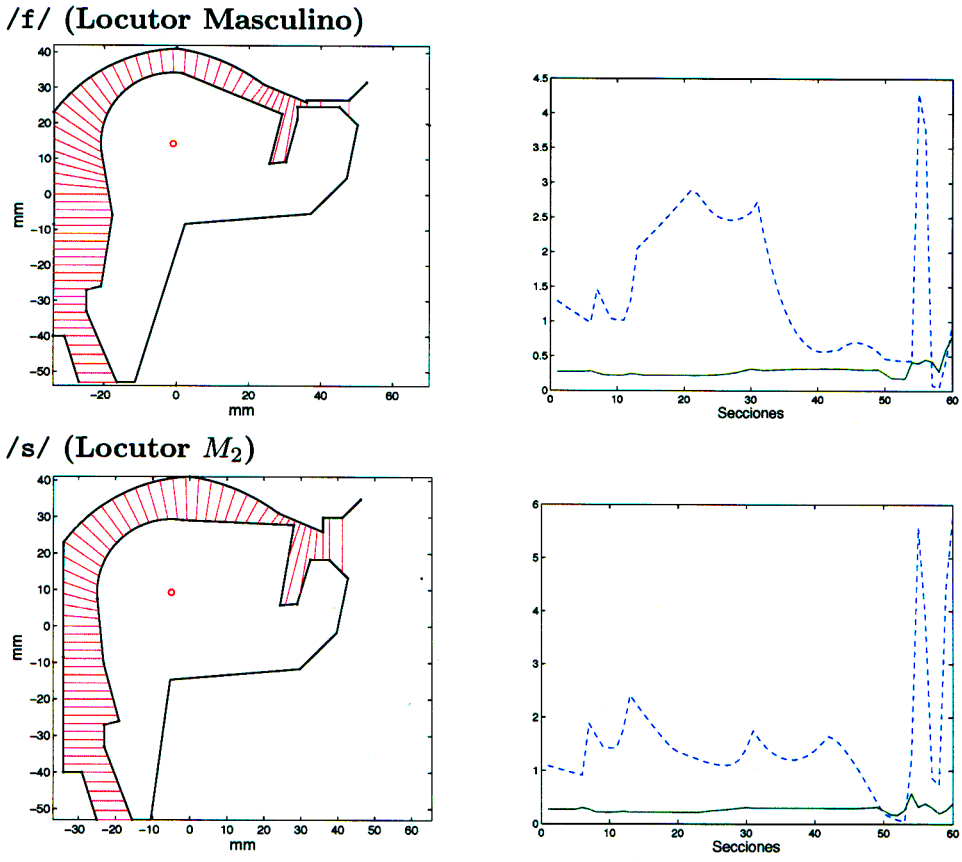


Figura 5.18: *Mejores configuraciones articulatorias recuperadas (consonantes /f/ y /s/). A la derecha de cada configuración medial se muestran las funciones respectivas de área (punteada, en cm²) y longitud (continua, en cm).*

En estas inversiones la actividad muscular también se muestra consistente. En la /s/, el GGm avanza la lengua y el UpI eleva el ápice. Por su parte, en la /f/, el OO acerca los labios, formando el tubo acústico anterior en cuyo análisis se ha basado la inversión de este fonema. El efecto de otros músculos resulta leve o vano.

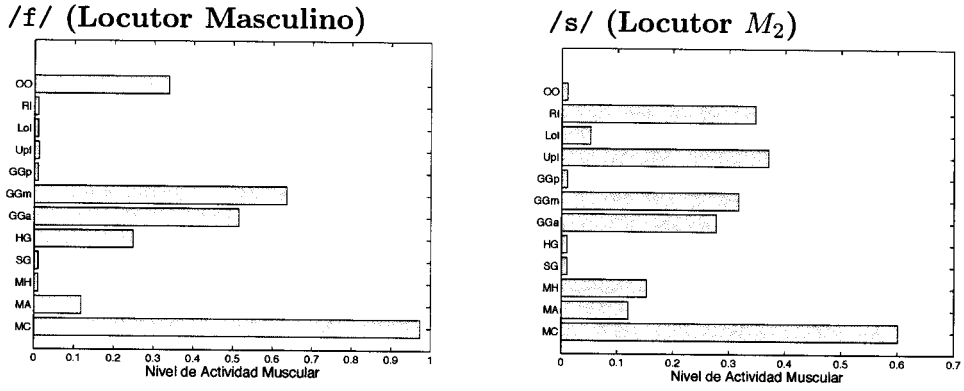


Figura 5.19: *Actividad muscular vinculada a las configuraciones articulatorias recuperadas (consonantes /f/ y /s/)*

	Resonancia	
Consonante /f/	P1	Error
Masculino	18.0000 17.4140	9.4028
Femenino	18.0000 17.4866	11.4566
Consonante /s/		
M_1	4.7273 4.8520	11.0672
M_2	6.0606 5.9800	7.6287
F_1	6.0533 6.3005	11.5634
F_2	6.2010 6.4132	10.5669

Cuadro 5.3: *Evaluación de la función objetivo con las mejores configuraciones recuperadas de la /f/ y la /s/.*

5.6. EVALUACIÓN SUBJETIVA DE LAS CONFIGURACIONES APRENDIDAS

La evaluación de una señal de voz respecto a naturalidad, calidad e inteligibilidad constituye un tópico de investigación por sí mismo [16, 40]. Sin embargo, para los fines del estudio, basta que la emisión sintética producida con los parámetros recuperados pueda ser reconocida correctamente por un humano, independientemente de los atributos subjetivos específicos de la señal objeto invertida. Las señales a evaluar se generaron con el sintetizador articulatorio del Capítulo 4, empleando los mejores vectores articulatorios recuperados en este capítulo, y como fuentes de excitación, la salida de la ESN en la sección 2.4 y el modelo de turbulencia de la sección 4.6.1. Naturalmente, la síntesis a partir de configuraciones femeninas requiere el ajuste de las dimensiones del modelo articulatorio y el uso de la red ESN para el locutor femenino prototipo.

La síntesis articulatoria de cada vocal implicó la producción de 4000 muestras, que según la frecuencia de discretización del modelo acústico, equivalen a una señal de 0.20 segundos. La Figura 5.20 exhibe los espectrogramas de las señales sintéticas derivadas de las mejores configuraciones articulatorias. En general los formantes se ubican apropiadamente; las diferencias proceden de la distorsión de frecuencia, y resultan inevitables para la síntesis en el dominio del tiempo.

En el caso de las consonantes, se sintetizaron las secuencias *ma*, *na*, *sa* y *fa*. La Figura 5.21 recopila las señales producidas con las mejores configuraciones. La vocal /a/ que sigue a la consonante proviene de la configuración invertida al locutor M_1. La síntesis de las secuencias se desarrolla en tres tramas. La primera corresponde exclusivamente a la consonante. Luego, una rutina integrada al sintetizador interpola las funciones de área y de longitud, con el fin de simular la transición desde la consonante hacia la vocal. La última trama agrupa las muestras de la /a/. Específicamente, en las nasales la primera trama tiene una longitud de 1000 muestras, la transición 500, y la vocal ocupa 4000 muestras. Como se mencionó en el Capítulo 4, y en la sección 5.5.4, el área velofaríngea se iguala a 0.5 cm^2 en la primera trama. Esta área disminuye durante la transición, hasta alcanzarse el cierre completo al inicio de la tercera trama. Por su parte, la síntesis de fricativas distribuye las muestras como sigue: 1000, 1000 y 2000. Además, por la afinidad entre la configuración de la /m/ y los fonemas /p/ y /b/ [78], también se generaron secuencias *pa* y *ba*, con la finalidad de explorar el alcance del modelo acústico. Recuérdese que las consonantes representan el tipo de señal más complicado para los sintetizadores articulatorios. Estos fonemas oclusivos se generan con una excitación impulsiva (fonema /p/), y cerrando el tracto oral por un lapso de tiempo, abriéndolo luego abruptamente (fonema /b/). En la /p/, el impulso ocupa 100 muestras, y las otras dos tramas, 700 y 3000, respectivamente. A su vez, la /b/ dispone 1000 muestras al comienzo, 500 en la transición, y 4000 para la vocal.

Con el propósito de precisar la calidad de las emisiones sintéticas, se constituyó un grupo de 8 evaluadores para pruebas de percepción. Con cada evaluador, una función ordena aleatoriamente las señales artificiales y las reproduce. Después de reproducir una señal, el programa solicita al evaluador que indique la vocal o secuencia que había oído. El programa contabilizaba un error si el evaluador ingresaba una categoría distinta a la ejecutada. Debe acotarse que los evaluadores ignoraban la proporción de señales en el corpus de prueba.

Las evaluaciones de vocales y consonantes se efectuaron por separado. Satisfactoriamente, el reconocimiento de vocales no incurrió en ningún error, lo cual confirma el éxito de la inversión desarrollada. Respecto a las consonantes, algunas secuencias fueron confundidas, como revela la matriz de confusión en el Cuadro 5.4. Para estos experimentos de percepción se sintetizaron cuatro instancias de las secuencias *ma*, *na*, *fa*, *sa*, *pa* y *ba*. Cada instancia se derivó a partir de las mejores configuraciones invertidas a los 4 locutores del corpus de señales objeto. Luego, como a cada evaluador se le han reproducido las 4 instancias de cada secuencia, se efectuaron 32 clasificaciones por secuencia.

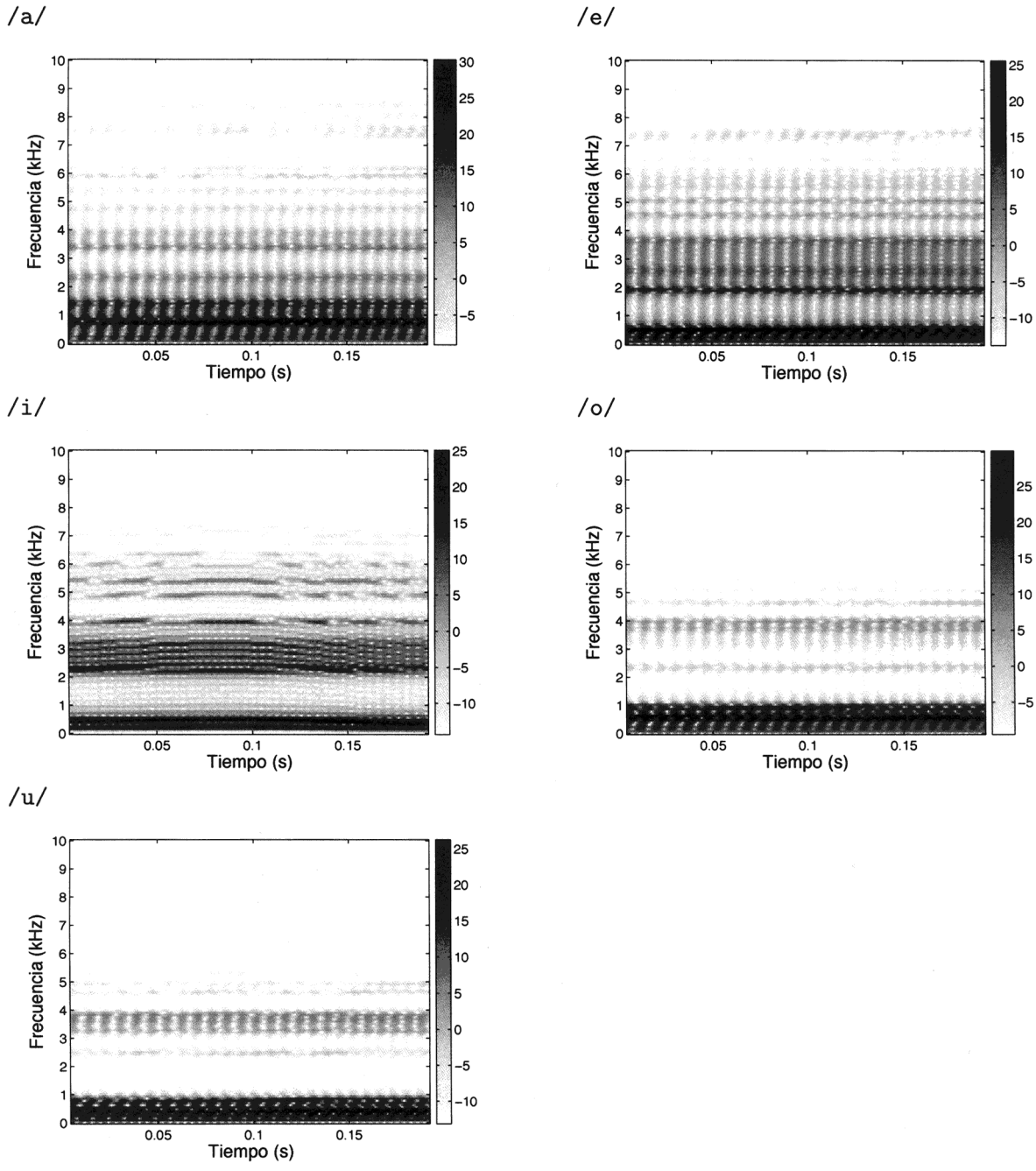


Figura 5.20: *Espectrogramas de algunas vocales sintetizadas*

	ma	*na*	*fa*	*sa*	*pa*	*ba*
ma	**0.4375**	0.4063	0.0000	0.0000	0.0313	0.1250
na	0.3125	**0.4688**	0.0000	0.0000	0.0000	0.2188
fa	0.0000	0.0000	**0.7813**	0.2188	0.0000	0.0000
sa	0.0000	0.0000	0.3125	**0.6875**	0.0000	0.0000
pa	0.0625	0.0000	0.0000	0.0000	**0.6250**	0.3125
ba	0.1250	0.0000	0.0000	0.0000	0.3125	**0.5625**

Cuadro 5.4: *Matriz de confusión en la identificación de consonantes.*

Por ejemplo, la primera fila de la matriz de confusión corresponde a las evaluaciones de las instancias de la secuencia *ma*. La tasa de 0.4375 significa que 14 señales se clasificaron correctamente como *ma*, mientras que 13 se identificaron, incorrectamente, como *na*. Además, 1 señal se clasificó como *pa*, y 4 como *ba*. No se presentaron confusiones de las instancias de *ma* con las secuencias fricativas *fa* y *sa*. Las otras filas de la matriz contienen las tasas de confusión correspondientes al resto de secuencias de consonantes sintetizadas.

A pesar de los errores, las secuencias identificadas correctamente siempre superan a las mal clasificadas. El Cuadro 5.4 también refleja que la mayor confusión se presenta distinguiendo entre los fonemas más próximos articulatoria y acústicamente. Particularmente problemática resulta la distinción entre las secuencias *ma* y *na*, lo que implica que el rango de frecuencias bajas considerado durante la inversión de estos fonemas resulta insuficiente. Recuérdese que la inversión excluyó información acústica sobre las transiciones, importante para el reconocimiento de las secuencias [57]. Por su parte, las tasas de 0.7813 y 0.6875 resultan muy buenas, considerando que el modelo de turbulencia de la sección 4.6.1 impone restricciones severas respecto a la realidad aeroacústica involucrada. En la actualidad, el conocimiento sobre el espectro, nivel, impedancia y distribución espacial de las señales fricativas no resulta completo ni claro [69]. Además, los espectrogramas exhiben perturbaciones en la transición, derivadas del bajo número de muestras en la trama; la única solución a este problema es el incremento de la frecuencia de muestreo, aunque con un mayor costo de procesamiento. Por su parte, las tasas de reconocimiento de las oclusivas, aunque un poco más bajas, resultan satisfactorias por cuanto en realidad el modelo acústico no se construyó con este tipo de fonemas como objetivo. En conclusión, estos resultados demuestran que el modelo acústico, a pesar de sus limitaciones, puede producir secuencias inteligibles con consonantes.

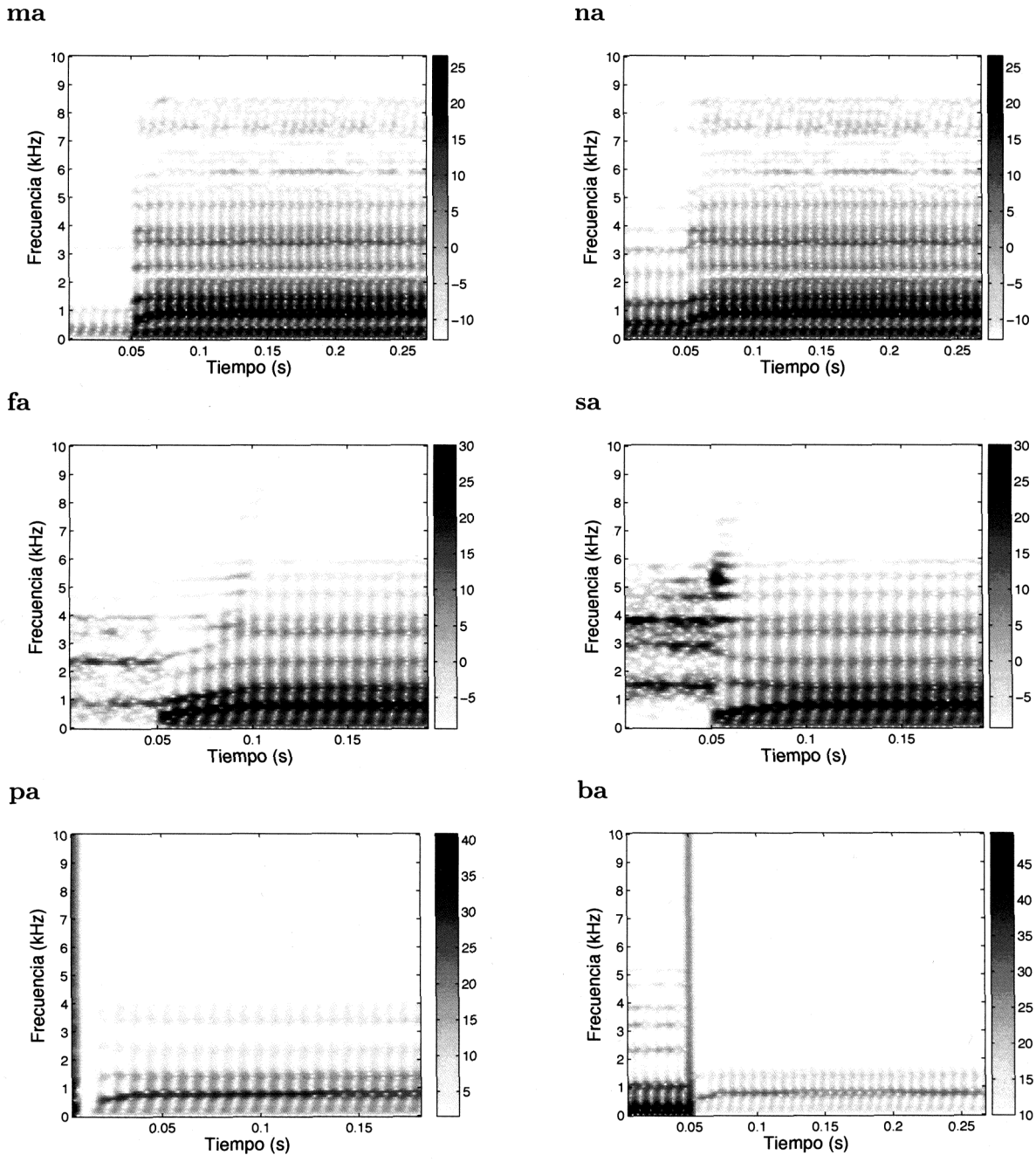


Figura 5.21: *Espectrogramas de algunas secuencias sintetizadas con consonantes*

6

Conclusiones y Recomendaciones

Como revelan las evaluaciones objetivas y subjetivas, todas las técnicas de aprendizaje empleadas abordaron satisfactoriamente el problema inverso, en sus distintas facetas. Así, se construyeron Redes con Estados de Eco capaces de aprender la señal de excitación glotal introduciendo automáticamente la variación de frecuencia jitter. Además, la red logró asociar una señal de entrada reguladora con la dinámica del DR para simular los cambios de amplitud de los pulsos glotales. Esta variabilidad en F0 contribuye con la naturalidad de las emisiones sintéticas. La identificación perfecta de las vocales por parte de los evaluadores indica que desaparecieron en buena medida los efectos de distorsión de las fuentes estrictamente regulares. Parte del éxito en el aprendizaje de la red proviene de la supresión de la fase cerrada en las señales de entrenamiento, lo cual, sin embargo, obliga a la reconstrucción de la salida de las redes antes de utilizarla con el modelo acústico. En este sentido, futuras investigaciones deben considerar la extensión de las ESN para abordar señales lentas, con segmentos relativamente invariables y largos. Las ESN con neuronas tipo *leaky-integrator* resultan prometedoras para este problema [30]. Nótese también que las señales de entrenamiento provienen de locutores prototipo extraídos de la literatura. En la aplicación de copia de un locutor, resultará necesario derivar la información sobre la excitación a partir de la propia señal objeto. Con este fin, puede recurrirse a herramientas como APARAT, una librería de MATLAB para la inversión del pulso glotal [1].

Por su parte, el modelado de las relaciones entre contracción muscular y cambio de la masa lingual en la configuración medial con un sistema de inferencia difuso TSK no planteó ninguna complicación a la inversión articulatoria. Por el contrario, simplificó la representación de la dinámica de la lengua, en torno al punto $B(x, y)$ del modelo articulatorio. A medida que surjan nuevos resultados respecto a la neurofisiología de los músculos supraglotales, la base de reglas puede modificarse fácilmente sin necesidad de alterar el funcionamiento del AGC. De esta forma, no sólo los músculos extrínsecos de la lengua se incorporarían al sistema difuso, sino todos los músculos del vector articulatorio.

A su vez, los Algoritmos Genéticos Continuos recuperaron exitosamente las configuraciones articulatorias asociadas a las señales objeto en el corpus. Las configuraciones revelan consistencia articulatoria, y se alcanzaron los umbrales de error con las funciones objetivo definidas. En el futuro, el proceso debe refinarse para evitar que la función objetivo dependa del tipo de fonema a invertir. Específicamente, se requiere una función objetivo suficientemente general, que no exija ningún tipo de modificación a priori. En una primera aproximación, para los fonemas estudiados aquí, debe distinguirse entre espectros sin ruido y con ruido, y en los primeros, detectar los picos de energía. Particularmente, la presencia de antiresonancias obstaculiza el análisis de las nasales. Aquí funcionó por la cantidad reducida de señales del corpus. Pero en otros estudios resultará necesario apelar a métodos más elaborados que la Predicción Lineal de orden variable. Por su parte, con las fricativas resulta imperioso trascender el análisis del tubo acústico anterior a la zona de constricción. Para ello se necesita un modelo aeroacústico que permita predecir con mayor propiedad la composición

espectral, a partir de una configuración articulatoria dada. En última instancia, se trata de ubicar las zonas espectrales con mayor contribución energética, y la forma en que la energía se distribuye en dichas regiones.

La inversión de otros fonemas como líquidas, oclusivas y glides, por ejemplo, exige cambios drásticos en el algoritmo, y en los modelos articulatorio y acústico. En primer lugar, la inversión no puede ser estática. Los experimentos desarrollados aquí no revelaron inconvenientes severos con el estatismo, por el cuidado en la formación del corpus. Pero en otros fonemas necesariamente habrá que recuperar una *secuencia* de eventos acústicos, y por ende la inversión debe planificarse desde una perspectiva dinámica. Estos modelos, si se generaliza la función objetivo como menciona el párrafo precedente, podrían invertir señales más complejas como palabras u oraciones. En tal caso, la señal objeto se dividiría en tramas aplicando el AGC en cada una. Naturalmente, para descartar cambios abruptos entre las configuraciones recuperadas, debe introducirse una métrica de distancia entre configuraciones de tramas vecinas, favoreciendo las trayectorias articulatorias que muestren un cambio más suave en los componentes del vector articulatorio.

En relación con el modelo articulatorio, éste proporcionó una satisfactoria reducción del espacio articulatorio, convirtiendo los 120 valores de las funciones de área y longitud combinadas en sólo 12 del vector articulatorio. Las evaluaciones del estudio superan las de investigaciones que emplean los 120 valores [9]. Sin embargo, se trata de un modelo desarrollado sobre la base de rectas y arcos regulares, por lo que con ciertas señales objeto el modelo recurre a cambios compensatorios, con el fin de reproducir las características acústicas del locutor. En este sentido, con un costo computacional notablemente elevado, algunas mejoras aplican:

1. El modelo debe emplear splines en su descripción geométrica, o extenderse a la variante tridimensional construida a partir de resonancias magnéticas. Ambas constituyen representaciones más flexibles y precisas de los tractos supraglotales, y también mucho más exigentes en términos de procesamiento computacional, al momento de derivar las funciones de área y longitud.

2. El AGC podría alterarse para abarcar también el ajuste automático de las dimensiones del modelo articulatorio, lo que en principio haría innecesaria la distinción entre señales objeto provenientes de locutores masculinos y femeninos.

3. La inversión de los fonemas con oclusión oral `/m/` y `/n/`, demuestra la conveniencia de un modelo más realista de los tejidos blandos. Básicamente, debe detectarse la colisión de las fronteras, y que el modelo reproduzca la deformación de los tejidos en contacto.

En general, las evaluaciones subjetivas ratifican los problemas de la síntesis articulatoria con las consonantes. Aunque la inversión satisfizo exitosamente, la síntesis revela carencias originadas en las simplificaciones y supuestos del modelo acústico. Las señales sintéticas de `/b/` y `/p/`, no invertidas, tampoco resultan distinguibles con toda claridad. Las mejoras a este modelo, sin embargo, dependen casi exclusivamente de los progresos en Acústica Teórica. Empero, una mejora que puede aplicar, con mayor capacidad de cómputo, es el incremento de la frecuencia de discretización para reducir la distorsión de frecuencia y la perturbación en las transiciones entre tramas durante la síntesis.

Finalmente, futuras investigaciones podrían beneficiarse de la separación en dos niveles del análisis aplicado en la inversión. Un primer nivel correspondería netamente al análisis acústico de las señales objeto (que en el AGC se representa con f_1), mientras el otro nivel se concentra en el vector articulatorio (asignado aquí, de forma restringida, a f_2), incorporando músculos adicionales y métricas más elaboradas, como por ejemplo, la asignación de costos a diversas actividades musculares, para modelar el mayor trabajo relativo de algunas contracciones, y para representar apropiadamente las relaciones de antagonismo entre músculos.

Bibliografía

[1] M. Airasa, H. Pulakka, T. Bäckström, y P. Aiku. A toolkit for voice inverse filtering and parametrisation. En *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005 - Eurospeech)*, páginas 2145–2148. Lisboa, Portugal, 2005.

[2] P. Badin y G. Fant. Notes on vocal tract computation. Informe técnico, STL-QPSR 2-3/1984, 1984.

[3] N. U. Baier. *Approximately Periodic Time Series and Nonlinear Structures.* Tesis Doctoral, École Polytechnique Fédérale de Lausanne, Lausanne, 2005.

[4] G. Bailly. Learning to speak: sensori-motor control of speech movements. *Speech Communication*, 22(2-3):251–267, 1998.

[5] R. Ball. Introduction to phonetics for students of english, french, german and spanish. `http://www.lang.soton.ac.uk/profiles/ball.htm`. University of Southampton. Visitado: 20 de Noviembre de 2005.

[6] P. Birkholz, D. Jackel, y B. J. Kröger. Construction and control of a three-dimensional vocal tract model. *IEEE 2006 International Conference on Acoustics, Speech and Signal Processing*, 2006.

[7] C. Blackburn y S. Young. A self-learning predictive model of articulator movements during speech production. *Journal of the Acoustical Society of America*, 107(3):1659–1670, Marzo 2000.

[8] P. Boersma. *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives.* W.H. Freeman and Company, The Hague, Netherlands, 1998. Primera edición.

[9] J. Brito. Genetic learning of vocal tract area functions for articulatory synthesis of spanish vowels. *Journal of Applied Soft Computing.* Aceptado para publicación.

[10] J. Brito. *Identificación de Señales Verbales en el Espacio de Fase Reconstruido.* Tesis de maestría. Universidad de Los Andes, Mérida, 2004.

[11] J. Brito y W. Rodríguez. Neural networks based speech signal classification with reconstructed dynamics features. En *Proceedings of the 2004 International Conference on Modelling and Simulation.* Valladolid, España, 2004.

[12] J. Brito y W. Rodríguez. Classification of spanish vowels and digits using neural networks. *WSEAS Transactions on Systems*, 4:921–924, 2005.

[13] J. Brito y W. Rodríguez. Recovering vocal tract area functions from spanish vowels using genetic algorithms. *WSEAS Transactions on Computers*, 4(12):1816–1823, 2005.

[14] J. Brito y W. Rodríguez. Multipopulation genetic learning of midsagittal articulatory models for speech synthesis. En *Proceedings of the 2006 IEEE International Conference on Granular Computing*. Atlanta, USA, 2006.

[15] M. A. Carreira-Perpiñán. *Continuous latent variable models for dimensionality reduction and sequential data reconstruction*. Tesis Doctoral, Department of Computer Science, University of Sheffield, Febrero 2001.

[16] D. G. Childers. *Speech Processing and Synthesis Toolboxes*. John Wiley and Sons, 2000.

[17] J. Dang y K. Honda. Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. *Journal of Phonetics*, 30:511–532, 2002.

[18] J. Dang y K. Honda. Construction and control of a physiological articulatory model. *Journal of the Acoustical Society of America*, 115:853–878, 2004.

[19] P. Denes y E. Pinson. *The Speech Chain: The Physics and Biology of Spoken Language*. W.H. Freeman and Company, New York, USA, 1993. Segunda edición.

[20] S. Dusan y L. Deng. Acoustic-to-articulatory inversion using dynamic and phonological constraints, 2000. 5th Speech Production Seminar.

[21] G. Fant. *Acoustic Theory of Speech Production*. Description and Analysis of Contemporary Standard Russian. Mouton, The Hague, The Netherlands, 1970.

[22] J. L. Flanagan. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, Berlin, 1972. Segunda edición.

[23] R. Halavati, S. B. Shouraki, y S. H. Zadeh. Recognition of human speech phonemes using a novel fuzzy approach. *Journal of Applied Soft Computing*, 2006. Doi:10.1016/j.asoc.2006.02.007 (En prensa).

[24] R. L. Haupt y S. E. Haupt. *Practical Genetic Algorithms*. Wiley-Interscience, 2004. Segunda edición.

[25] S. Haykin. *Neural Networks: A comprehensive foundation*. McMillan Publishing Company, 1994. Primera edición.

[26] F. Herrera, M. Lozano, y J. L. Verdegay. Tackling real-coded genetic algorithms: Operators and tools for behavioural analysis. *Artificial Intelligence Review*, 12(4):265–319, 1998.

[27] W. Hess. Artikulatorische und akustische phonetik: Die menschlichen sprech-organe. `http://www.ikp.uni-bonn.de/dt/lehre/materialien/aap/aap_1f.pdf`, 2005. Institut für Kommunikationsforschung und Phonetik, Universität Bonn. Visitado: 12 de Septiembre de 2005.

[28] K. Ishizaka y J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Syst. Tech. Jour.*, 51:1233–1268, 1972.

[29] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Informe técnico, German National Research Center for Information Technology, 2001.

[30] H. Jaeger. A tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the "echo state network" approach. Informe técnico, Fraunhofer Institute for Autonomous Intelligent Systems, 2005.

[31] J. K. Kelly y C. C. Lochbaum. Speech synthesis. En *Proceedings of the 4th International Congress of Acoustics.* Morgan Kaufmann, San Diego, USA, 1962.

[32] J. Kelso, E. Saltzman, y B. Tuller. The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, 14:29–59, 1986.

[33] D. H. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

[34] B. J. Kröger. Comunicación privada, 2006.

[35] B. J. Kröger, R. Winkler, C. Mooshammer, y B. Pompino-Marschall. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. En *Proceedings of the 5th Seminar on Speech Production.* Bavaria, Germany, 2000.

[36] M. J. Lambeth y M. J. Kushmerick. A computational model for glycogenolysis in skeletal muscle. *Annals of Biomedical Engineering*, 30:808–827, 2002.

[37] J. Larar, J. Scroeter, y M. Sondhi. Vector quantisation of the articulatory space. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(12):1812–1818, Diciembre 1988.

[38] A. S. Leonov y V. Sorokin. Inverse problem for the vocal tract: Identification of control forces from articulatory movements. *Pattern Recognition and Image Analysis*, 10:110–126, 2000.

[39] A. S. Leonov y V. Sorokin. Optimality criteria in inverse problems for tongue-jaw. En *Euro-Speech 2003.* Genova, 2003.

[40] F. F. Li y T. J. Cox. A neural network model for speech intelligibility quantification. *Journal of Applied Soft Computing*, 2005. Doi:10.1016/j.asoc.2005.05.002 (En prensa).

[41] Q. Lin. *Speech Production Theory and Articulatory Speech Synthesis.* Tesis Doctoral, Royal Institute of Technology, Stockholm, Sweden, 1990.

[42] S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1:199–229, 1982.

[43] S. Maeda. The role of the sinus cavities in the production of the nasal vowels. En *IEEE 1982 International Conference on Acoustics, Speech and Signal Processing*, páginas 911–914. 1982.

[44] S. Maeda y K. Honda. From emg to formant patterns of vowels: The implication of vowel spaces. *J. Phonetica*, 51:17–29, 1994.

[45] R. McGowan. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 14(1):19–48, Febrero 1994.

[46] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082, 1973.

[47] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag, New York, 1994. Segunda edición.

[48] T. Mitchell. *Machine Learning.* McGraw-Hill, United States, 1997.

[49] H. Mühlenbein y D. Schlierkamp-Voosen. Predictive models for the breeder genetic algorithms. *Evolutionary Computation*, 1:25–49, 1993.

[50] E. Obediente Sosa. *Fonética y Fonología.* Consejo de Publicaciones de la Universidad de Los Andes, 2001. Tercera edición.

[51] S. Ouni y Y. Laprie. Articulatory space modeling using hypercubes for the acoustic-to-articulatory inversion. *Journal of Acoustical Society of America*, 2003.

[52] A. Oyama, S. Obayashi, y T.Nakamura. Real-coded adaptive range genetic algorithm applied to transonic wing optimization. En H.-P. S. Marc Schoenauer, Kalyanmoy Deb, Günter Rudolph, Xin Yao, Evelyne Lutton, Juan Julian Merelo, editor, *Parallel Problem Solving from Nature - PPSN VI 6th International Conference.* Springer Verlag, Paris, France, 16-20 2000.

[53] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, y S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the Acoustical Society of America*, 92(2):688–700, 1992.

[54] H. Ploner-Bernard. Speech synthesis by articulatory models. Informe técnico, Graz University of Technology, 2003.

[55] J. B. Porta. *Natural Magick.* Nuvision Publications, 2005.

[56] M. R. Portnoff. *A Quasi-One Dimensional Digital Simulation for the Time-varying Vocal Tract.* Tesis de maestría, MIT, 1973.

[57] T. Quatieri. *Discrete-time speech signal processing.* Prentice-Hall, New Jersey, USA, 2002.

[58] L. R. Rabiner y R. W. Schafer. *Digital Processing of Speech Signals.* Signal Processing. Prentice-Hall, New Jersey, 1978.

[59] M. Rahim, C. Goodyear, W. Kleijn, J. Schroeter, y M. Sondhi. On the use of neural networks in articulatory speech synthesis. *Journal of the Acoustical Society of America*, 93(2):1109–1121, Febrero 1993.

[60] M. G. Rahim y C. C. Goodyear. Estimation of vocal tract filter parameter using a neural net. *Speech Communication*, 9, 1990.

[61] W. Rodríguez. *Similarity of Dynamical Systems.* Tesis Doctoral, University of South Florida, 1998.

[62] W. Rodríguez, H.-N. Teodorescu, F. Grigoras, A. Kandel, y H. Bunke. A fuzzy information space approach to speech signal non-linear analysis. *International Journal of Intelligent Systems*, 15(4):343–363, 2000.

[63] P. Rubin, T. Baer, y P. Mermelstein. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70:321–328, 1981.

[64] K.-I. Sakakibara y H. Imagawa. A many-parameter model of laryngeal flow with ventricular resonance and supraglottal vibration. En *Proceedings of Forum Acusticum 2005*. Budapest, Hungary, 2005.

[65] E. Saltzman. Task-dynamic coordination of the speech articulators: A preliminary model. *Experimental Brain Research*, 15, 1986.

[66] R. A. Scarborough. Lexical confusability and degree of coarticulation. En *Proceedings of the 29th Meeting of the Berkeley Linguistics Society*. Berkeley, United States, 2003.

[67] J. Schroeter y M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2, 1994.

[68] M. Senthil Arumugama, M. Rao, y R. Palaniappan. New hybrid genetic operators for real coded genetic algorithm to compute optimal control of a class of hybrid systems. *Journal of Applied Soft Computing*, 6(1):38–52, 2005.

[69] D. J. Sinder. *Speech synthesis using an aeroacoustic fricative model.* Tesis Doctoral, Graduate School of Electrical and Computer Engineering, The State University of New Jersey, 1999.

[70] M. Sondhi. Resonances of a bent vocal tract. *Journal of the Acoustical Society of America*, 79:1113–1116, 1986.

[71] M. Sondhi y J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7):955–967, 1994.

[72] V. Sorokin. Inverse problem for fricatives. *Speech Communication*, 14:249–262, 1994.

[73] V. Sorokin. Articulatory-to-acoustic mapping for inverse problem. *Speech Communication*, 19:105–118, 1996.

[74] V. Sorokin. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30:55–74, 2000.

[75] V. Sorokin. Some coding properties for speech. *Speech Communication*, 40:409–423, 2003.

[76] V. Sorokin, A. S. Leonov, I. Makarov, y A. Tsyplikhin. Speech inversion and resynthesis. En *Interspeech 2005*. Lisboa, Portugal, 2005.

[77] V. Sorokin, V. Olshansky, y L. Kozhanov. Internal model in articulatory control: Evidence from speaking without larynx. *Speech Communication*, 25:249–268, 1998.

[78] K. Stevens. *Acoustic Phonetics*. Current studies in linguistic. The MIT Press, Massachusetts, 1998.

[79] M. Stone, E. Davis, A. Douglas, M.Nessaiver, R. Gullapalli, W. Levine, y A. Lundberg. Modeling the motion of the internal tongue from tagged cine-mri images. *Journal of the Acoustical Society of America*, 109:2974–2982, 2001.

[80] B. H. Story y I. R. Titze. Voice simulation with a body cover-cover model of the vocal folds. *Journal of Acoustical Society of America*, 97(2):1249–1260, 1995.

[81] M. Sugeno y G. T. Kang. Fuzzy modeling and control of multilayer incinerator. *Fuzzy Sets and Systems*, 18:329–346, 1986.

[82] T. Takagi y M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15:116–132, 1985.

[83] The UCLA Phonetics Laboratory. *Dissection of the Speech Production Mechanism.* `http://www.linguistics.ucla.edu/people/ladefoge/manual.htm`, United States, 2002. Visitado: 17 de Abril de 2006.

[84] I. Titze, R. Baken, y H. Herzel. *Vocal Fold Physiology: New Frontier in Basic Science*, capítulo Evidence of chaos in vocal fold vibration, páginas 143–188. Whurr Books Publishers, 1993.

[85] D. Ünay. *Analysis of Tongue Motion using Tagged CINE-MRI.* Tesis de maestría, Bogazici University, July 2001.

[86] H. Wakita y G. Fant. Toward a better vocal tract model. Informe técnico, STL-QPSR 1/1978, 1978.

[87] D. Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4:65–85, 1994.

[88] D. Whitley y T. Starkweather. Genitor 2: a distributed genetic algorithm. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:189–214, 1990.

[89] L. F. Wilde. *Analysis and synthesis of fricative consonants.* Tesis Doctoral, MIT, Cambridge, 1995.

[90] O. Wolkenhauer. *Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis.* Wiley-Interscience, New York, 2001.

[91] A. H. Wright. *Foundations of Genetic Algorithms*, capítulo Genetic algorithms for real parameter optimization, páginas 205–218. Morgan Kaufman, San Mateo, 1991.

[92] C. Wu. *Articulatory Speech Synthesizer.* Tesis Doctoral, University of Florida, Florida, United States, 1996.

[93] Q. Yan, S. Vaseghi, E. Zavarehei, y B. Milner. Formant-tracking linear prediction models for speech processing in noisy environments. En *Interspeech 2005.* Lisboa, Portugal, 2005.

[94] H. Yehia y F. Itakura. A method to combine acoustic and morphologial constraints in the speech production inverse problem. *Speech Communication*, 18, 1996.

[95] Y. Zhang y J. Jiang. Chaotic vibrations of a vocal fold model with an unilateral polyp. *Journal of the Acoustical Society of America*, 115:1266–1269, 2004.