Departamento

Arquitectura y Tecnología de Sistemas Informáticos

Facultad de Informática

Desarrollo de algoritmos basados en filtrado adaptativo y su aplicación en el estudio de la fonética acústica española

Autor

Jesús Bobadilla Sancho
Licenciado en Informática

Director

Pedro Gómez Vilda
Catedrático de Universidad

1998

Deseo agradecer los consejos, ayuda y orientación que siempre me ha proporcionado Pedro Gómez en la dirección de esta tesis. También quisiera reconocer la importancia que los trabajos e investigaciones desarrollados con Victoria Rodellar y Mercedes Pérez han tenido en mi carrera.

Por último, quiero dar las gracias a aquellas personas que han compartido conmigo todo lo bueno y lo malo que supone realizar unos estudios tan largos; especialmente en el caso de Pilar, a quién dedico este libro.

RESUMEN

La hipótesis en la que se basa el desarrollo de esta tesis, se centra en la suposición de que partiendo del método de predicción lineal, es posible idear algoritmos de tratamiento de señal que permitan obtener una buena estimación de características espectrales significativas de la voz, especialmente en la detección de los formantes que se producen en el habla. Estos algoritmos ayudarían a construir un catálogo analítico de los principales sonidos del español, con el objetivo de complementar los estudios realizados hasta el momento en el campo de la fonética acústica.

Parte de la complejidad que presenta esta tesis doctoral, viene dada por la naturaleza multidisciplinar de las materias que aborda. La correcta determinación de diversas características espectrales del habla, requiere un amplio conocimiento de los fundamentos del tratamiento de la señal de voz y de la fonética del idioma escogido. También resulta necesario poseer nociones adecuadas de todas las áreas relacionadas con el tratamiento de la voz, con el fin de enfocar los estudios partiendo de una visión global del campo seleccionado.

Las investigaciones desarrolladas en este trabajo se han dividido en dos bloques fundamentales: tratamiento de señal y fonética acústica. En el apartado de tratamiento de señal, se ha validado la hipótesis inicial. La obtención de los formantes del habla se ha basado en el método de predicción lineal, haciéndose una búsqueda de polos fuera de la zona habitual (el círculo unidad). La decisión de trabajar con funciones espectrales suavizadas ha resultado muy adecuada para la estimación de los formantes de voz. Partiendo de estas funciones espectrales se han ideado diferentes etapas que van detectando y resaltando los formantes del habla haciendo uso de transformaciones no lineales basadas en métodos algorítmicos.

En el bloque reservado para las investigaciones en fonética acústica española, se aportan mapas tridimensionales de sonidos vocálicos que sirven como modelo para la extensión de las frecuentes clasificaciones bidimensionales que se utilizan en las publicaciones especializadas de fonética acústica. El empleo de una tercera dimensión permite complementar la información tradicional usada en las representaciones vocálicas. Así mismo se aportan trabajos que estudian la evolución de los formantes en situaciones de coarticulación. Estos trabajos se pueden considerar como una referencia innovadora para el desarrollo de investigaciones más elaboradas que se basen en los métodos y herramientas originales empleados en la tesis.

En esta tesis se ofrece abundante y variado material en forma de espectros típicos, generalización de la evolución de los formantes, planos de situación de vocales, etc. Estos datos y resultados, junto a la metodología y herramientas informáticas empleados, pueden servir de base para la creación de aplicaciones que actúen sobre distintas áreas del tratamiento de la voz, tales como la enseñanza asistida de idiomas, logopedia, reconocimiento y síntesis del habla, detección de discapacidades, modelizaciones acústicas basadas en la fonética, etc.

ABSTRACT

This thesis is based on the following hypothesis: using the linear prediction method, it is possible to devise signal processing algorithms which obtain a good estimation of significant spectral characteristics of the voice, specially the formants of the speech. These algorithms would help to obtain an analytical catalogue of the main sounds of the Spanish Language, and therefore complement the current studies in the acoustic/phonetics area.

Most of the complexity of this doctoral thesis comes from the different subjects covered by the speech processing area. The correct determination of diverse spectral characteristics in the speech, requires a deep knowledge in speech signal processing and the phonetics of the chosen language. In addition, it is necessary to incorporate a suitable background of all the subjects closely connected with speech processing.

The research carried out in this work has been classified in two main areas: signal processing and acoustic phonetics. In the signal processing field, the initial hypothesis has been validated. The linear prediction method has been used to get the speech formants, searching the poles outside the usual zone (the unit circle). Working with smoothed spectral functions has been very suitable to fix the speech formants. Starting from these spectral functions, different stages have been developed in order to detect and emphasize the speech formants using nonlinear transformations based on algorithmic methods.

With respect to the acoustic phonetics of Spanish, three-dimensional maps of vocalic sounds have been obtained. These maps can serve as a model to extend the two-dimensional classifications used in specialized publications of acoustic phonetics. The third dimension allows to complement the traditional information used in the vocalic representations. The formant evolution in "vowel-consonant-vowel" situations has been studied too. This work may be considered as a reference for future research based on the original methods and tools developed.

Finally, abundant and varied material is offered in form of typical time-frequency representations, formant evolutions, two-dimensional and three-dimensional maps of vowels, etc. These data and results, the methodology, and the computing tools developed, can serve as a base to create applications related with different speech processing areas, such as computer assisted language learning, recognition and speech synthesis, acoustic modeling based on the phonetics, etc.

INDICE

1.- Introducción 1

2.- Objetivos 5

3.- Estado del arte
- 3.1. Introducción 7
- 3.2. Métodos y algoritmos básicos para la obtención de características de la señal de voz
  - 3.2.1. Análisis en el dominio del tiempo 9
  - 3.2.2. Análisis en el dominio de la frecuencia 13
- 3.3. Fonética acústica española 44
- 3.4. Fonética acústica. Líneas de investigación 52

4.- Caracterización espectral de sonidos
- 4.1. Introducción 66
- 4.2. Extracción de formantes en segmentos vocálicos
  - 4.2.1. Resumen 67
  - 4.2.2. Introducción 67
  - 4.2.3. Predicción lineal 71
  - 4.2.4. Estrategias para la obtención de formantes 74
  - 4.2.5. Desarrollos preliminares 75
  - 4.2.6. Conceptos básicos del método 76
  - 4.2.7. Primera aproximación al algoritmo 74
  - 4.2.8. Segunda aproximación al algoritmo 84
  - 4.2.9. Método de obtención de formantes 88
  - 4.2.10. Suavizado de la señal 90
  - 4.2.11. Resultados 91
  - 4.2.12. Conclusiones 93
- 4.3. Determinación de características espectrales en sonidos sonoros
  - 4.3.1. Resumen 95
  - 4.3.2. Introducción 95
  - 4.3.3. Algoritmos y funciones ideados 97
  - 4.3.4. Evolución de los espectros según las funciones empleadas 104
  - 4.3.5. Elección de la escala de visualización de los espectros 112
  - 4.3.6. Resultados 114
  - 4.3.7. Conclusiones 118
- 4.4. Determinación de características espectrales en sonidos sordos
  - 4.4.1. Resumen 119
  - 4.4.2. Introducción 119
  - 4.4.3. Detección de la sonoridad/sordez 120
  - 4.4.4. Funciones espectrales desarrolladas 123
  - 4.4.5. Espectros de frases obtenidos con las funciones propuestas 135
  - 4.4.6. Conclusiones 139
- 4.5. Conclusiones 140

5.- Análisis de la situación y evolución de formantes en sonidos de la lengua española
5.1. Resumen 141
5.2. Introducción 141
5.3. Sonidos vocálicos en un sólo hablante
5.3.1. Introducción 142
5.3.2. Desarrollo 143
5.3.3. Conclusiones 158
5.4. Sonidos no vocálicos en un sólo hablante
5.4.1. Introducción 159
5.4.2. Desarrollo 160
5.4.3. Conclusiones 174
5.5. Sonidos vocálicos en varios hablantes
5.5.1. Introducción 175
5.5.2. Metodología empleada 175
5.5.3. Resultados obtenidos en cada uno de los hablantes 177
5.5.4. Estadísticos relevantes 197
5.5.5. Conclusiones 206
5.6. Sonidos no vocálicos en varios hablantes
5.6.1. Introducción 207
5.6.2. Consonantes bilabiales 209
5.6.3. Consonantes dentales/interdentales 210
5.6.4. Consonantes velares/palatales 211
5.6.5. Consonantes fricativas/africada 212
5.6.6. Consonantes líquidas 214
5.6.7. Conclusiones 215
5.7. Conclusiones 216

6.- Conclusiones
6.1. Conclusiones generales 217
6.2. Aportaciones originales 220
6.3. Ampliaciones propuestas 221

7.- Bibliografía 222

APÉNDICE - Espectros de evolución de formantes
A.- Oclusivas sordas
B.- Fricativas sonoras
C.- Nasales
D.- Laterales/vibrantes
E.- Fricativas sordas / africada

LISTA DE FIGURAS

CAPÍTULO 3

3.1	Mecanismo de establecimiento de ventanas sobre una señal	9
3.2	Funciones de energía, cruces por cero y máximos en la palabra 'hipotenusa'	12
3.3	Ejemplo de descomposición de una señal compleja en sumatorio de señales simples (continuación)	13
3.4	Ejemplo de descomposición de una señal compleja en sumatorio de señales simples (continuación)	14
3.5	Ejemplo de descomposición de una señal compleja en sumatorio de señales simples (continuación)	15
3.6	Señal muestreada, con indicaciones del significado de m(k), T y kT	16
3.7	Comparación de funciones de diferentes frecuencias con un bloque de la señal que se pretende descomponer	20
3.8	Señal a descomponer y resultado de compararla con senos de diferentes frecuencias	21
3.9	Señal a descomponer y resultado de compararla con senos de diferentes frecuencias (continuación)	22
3.10	Ejemplos de los sonidos vocálicos orales y nasales	45
3.11	Ejemplos de sonidos oclusivos y de fricativos sonoros	47
3.12	Ejemplos de los sonidos nasales	49
3.13	Ejemplos de sonidos fricativos y africado	50
3.14	Ejemplos de sonidos líquidos	51

CAPÍTULO 4

4.1	Posiciones medias de los tres primeros formantes de las vocales españolas	68
4.2	Posiciones de los tres primeros formantes de las vocales castellanas en un hablante tomado como ejemplo	68
4.3	Niveles usuales empleados en el reconocimiento del habla, cargando la importancia del proceso en el método de reconocimiento o en el cálculo de características espectrales de nivel medio	70
4.4	Predicción de los datos adicionales de una señal aplicando LPC	71
4.5	Evolución de las funciones espectrales obtenidas aplicando diferentes radios a la función de transferencia	73
4.6	Ejemplo de evolución a lo largo del tiempo de los picos en las funciones espectrales	75
4.7	Ejemplo de evolución de las funciones espectrales entre r=0.8 y r=1 en un instante de tiempo de la vocal 'o'	77
4.8	Ejemplo de evolución de las funciones espectrales entre r=0.85 y r=0.95 en un instante de tiempo de 'o'	78
4.9	Ejemplo de función espectral evaluada en r=0.9 de las 5 vocales castellanas	79
4.10	Función espectral evaluada en r=0.9 en un intervalo de tiempo de una señal de voz correspondiente a 'o'	80
4.11	Ejemplo de función espectral evaluada en r=0.9 de las 5 vocales castellanas	81
4.12	Vocal 'i' en la que aparece un máximo que no se corresponde con un formante	82
4.13	Vocal 'a' en la que los máximos pasan de representar formantes de forma directa a no hacerlo	83
4.14	Caso de una vocal 'a' en la que resulta difícil determinar sus formantes al duplicarse un máximo en bajas frecuencias	83
4.15	División en secciones que se obtiene al hallar el mínimo global de la función espectral	85
4.16	Ejemplo en el que se ignoran máximos de la función derivada segunda que no deben ser tomados como formantes	86
4.17	Máximos tomados como formantes después de rebajar el valor de la media de la función espectral	87
4.18	Posiciones de los formantes obtenidos en vocales pronunciadas por dos hablantes	90
4.19	Espectros de vocales pronunciadas por diferentes hablantes	92
4.20	Formantes hallados en los espectros de la figura anterior	93
4.21	Polos, derivada segunda y función obtenida ponderando las anteriores	98

4.22	Polos, derivada segunda, función ponderada y función ponderada realzada obtenida a partir de las anteriores	99
4.23	Función espectral obtenida ponderando según las medias de la función espectral de minimización de error	100
4.24	Función de transformación para la obtención de la función espectral ponderada según las medias. Función de aplanamiento de los valores espectrales por debajo de la media aritmética de la señal de voz en tiempo	101
4.25	Fusión de las funciones obtenidas para la obtención de un espectro representativo y función final con los formantes realzados.	103
4.26	Espectro del triptongo 'ioi' usando como función la derivada segunda. Se presentan distintos puntos de vista de la misma figura	105
4.27	Espectro del triptongo 'ioi' usando la función ponderación	106
4.28	Espectro del triptongo 'ioi' obtenido usando la función espectral ponderada a partir de las medias de la función de error	107
4.29	Espectro del triptongo 'ioi' obtenido usando la función final	108
4.30	Espectros del sonido 'ieaou' usando diferentes funciones para su obtención	109
4.31	Espectros de los sonidos 'eme ene eɳe' usando diferentes funciones para su obtención	111
4.32	Ejemplos de espectros usando distintas escalas de colores	113
4.33	Sonidos 'ere erre ele eʎe' y 'imi eme ama omo umu' empleando los espectros propuestos y sus correspondientes LPC	114
4.34	Sonidos 'epe ete eke' y 'eβe ed.e eɣe' empleando los espectros propuestos y sus correspondientes LPC	115
4.35	Espectros finales utilizando el método propuesto	117
4.36	Características de la señal de voz para la detección de la sordez-sonoridad	121
4.37	Espectro LPC tradicional de la secuencia "ese efe etʃe exe eθe"	123
4.38	Polos y derivada segunda de cuatro instantes espectrales de diferentes sonidos sordos	124
4.39	Polos, función ponderación y función de aproximación de cuatro instantes espectrales de varios sonidos sordos	126
4.40	Realizaciones espectrales de silencios	127
4.41	Resultado de la aplicación del factor de intensidad sobre un espectro de voz	128
4.42	Espectros hallados utilizando la función de polos (gráfico superior) y la función de aproximación	129
4.43	Resultado de aplicar un filtro paso bajo a las funciones básicas	130
4.44	Resultados variando la anchura del filtro paso bajo sobre espectros hallados con la función de polos (2 primeros casos) y de aproximación (2 últimos)	131
4.45	Espectros de sonidos sordos articulados con diferentes vocales	132
4.46	Espectros LPC clásicos correspondientes a los de la figura 4.45	133
4.47	Características espectrales de las oclusivas sordas	134
4.48	Espectros generados a partir de la frase "cinco seis siete ocho"	136
4.49	Espectros generados a partir de la frase "os deseo felices fiestas"	137
4.50	Espectros generados a partir de la frase "se siente asfixiado".	138

CAPÍTULO 5

5.1	Espectro de voz de la secuencia 'i e a o u', junto a varios instantes espectrales de referencia	144
5.2	Posiciones de los tres primeros formantes de vocales aisladas confrontadas entre sí	145
5.3	Espectros de voz correspondientes al grupo nasal ('imi ini iɳi', 'eme ene eɳe', 'ama ana aɳa', 'omo ono oɳo', 'umu unu uɳu')	146
5.4	Posiciones de los tres primeros formantes de vocales junto a nasal confrontadas entre sí	147
5.5	Espectros de voz correspondientes al grupo oclusivo sordo ('ipi iti iki', 'epe ete eke', 'apa ata aka', 'opo oto oko', 'upu utu uku')	148

5.6	Posiciones de los tres primeros formantes de vocales junto a oclusivas sordas confrontadas entre sí	149
5.7	Espectros de voz correspondientes al grupo oclusivo sonoro ('iβi id.i iγi', 'eβe ed.e eγe', 'aβa ad.a aγa', 'oβo od.o oγo', 'uβu ud.u uγu')	150
5.8	Posiciones de los tres primeros formantes de vocales junto a oclusivas sonoras confrontadas entre sí	151
5.9	Espectros de voz correspondientes al grupo fricativo-africada ('ifi iθi isi ixi itʃi', 'efe eθe ese exe etʃe', 'afa aθa asa axa atʃa', 'ofo oθo oso oxo otʃo', 'ufu uθu usu uxu utʃu')	152
5.10	Posiciones de los tres primeros formantes de vocales junto a fricativas-africada confrontadas entre sí	153
5.11	Espectros de voz correspondientes al grupo laterales-vibrantes ('ili iri iλi irri' , 'ele ere eλe erre', 'ala ara aλa arra', 'olo oro oλo orro', 'ulu uru uλu urru')	154
5.12	Posiciones de los tres primeros formantes de vocales ante laterales-vibrantes confrontadas entre sí	155
5.13	Posiciones de los tres primeros formantes de vocales junto a consonantes, con separaciones atendiendo al modo de articulación	156
5.14	Posiciones de los tres primeros formantes de vocales junto a consonantes	156
5.15	Espectros de voz de las palabras 'patético', 'ocupa', 'picota'	157
5.16	Evolución de los formantes en las palabras 'patético', 'picota' y 'ocupa'	158
5.17	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido nasal [m]	161
5.18	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido nasal [n]	162
5.19	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido nasal [ɳ]	163
5.20	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido oclusivo sordo [p]	164
5.21	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido oclusivo sordo [t]	165
5.22	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido oclusivo sordo [k]	166
5.23	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido fricativo sonoro [β]	166
5.24	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido fricativo sonoro [d.]	167
5.25	Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido fricativo sonoro [γ]	167
5.26	Evolución de F2 en las vocales situadas ante consonantes de tipo nasal , oclusivo sordo y fricativo sonoro	168
5.27	Evolución de F3 en las vocales situadas ante consonantes de tipo nasal, oclusivo sordo y fricativo sonoro	169
5.28	Evolución de F2 en las vocales situadas ante consonantes de tipo bilabial, dental-interdental y palatal-velar	170
5.29	Evolución de F3 en las vocales situadas ante consonantes de tipo bilabial, dental-interdental y palatal-velar	171
5.30	Espectros típicos del sonido 'epe' con valores (en Hercios) de la posición y evolución de los formantes segundo y tercero	172
5.31	Espectros típicos del sonido 'eke' con valores (en Hercios) de la posición y evolución de los formantes segundo y tercero	173
5.32	Ejemplo de espectros sobre los que se obtienen los valores de los formantes	177
5.33	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales aisladas	178
5.34	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales adyacentes a consonantes bilabiales	179
5.35	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales adyacentes a consonantes dentales/interdentales	180
5.36	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales adyacentes a consonantes velares/palatales	180
5.37	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales aisladas, vocales adyacentes a consonantes bilabiales, a consonantes dentales/interdentales, a consonantes velares/palatales	181
5.38	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano	182
5.39	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos para un hablante femenino en vocales aisladas, vocales adyacentes a consonantes bilabiales, a consonantes dentales/interdentales, a consonantes velares/palatales	183
5.40	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando un hablante femenino	184

5.41	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos para un hablante femenino en vocales aisladas, vocales adyacentes a consonantes bilabiales, a consonantes dentales/interdentales, a consonantes velares/palatales	185
5.42	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando un hablante femenino	185
5.43	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos para un hablante masculino en vocales aisladas, vocales adyacentes a consonantes bilabiales, a consonantes dentales/interdentales, a consonantes velares/palatales	186
5.44	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando un hablante masculino	187
5.45	Espectros origen de los valores obtenidos en los formantes de la vocal 'a' en la figura 5.43. Espectro superior: 'apa aβa ama'. Espectro inferior: 'aka aɣa aŋa'. Espectro de la derecha: 'ata ad.a ana'	188
5.46	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en dos hablantes masculinos sobre vocales aisladas, vocales adyacentes a consonantes bilabiales, a consonantes dentales/interdentales, a consonantes velares/palatales	189
5.47	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando dos hablantes masculinos	189
5.48	Posiciones de los formantes obtenidos en los dos hablantes masculinos	190
5.49	Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en dos hablantes femeninos sobre vocales aisladas, vocales adyacentes a consonantes bilabiales, a consonantes dentales/interdentales , a consonantes velares/palatales	191
5.50	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando dos hablantes femeninos	192
5.51	Posiciones de los formantes obtenidos en los dos hablantes femeninos	192
5.52	Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando a los 4 hablantes (2 masculinos y 2 femeninos)	193
5.53	Posiciones de los formantes obtenidos empleando los 4 hablantes (2 masculinos y 2 femeninos)	194
5.54	Planos F1-F2 correspondientes a las figuras 5.52 y 5.53 presentados como nube de puntos y por superficies transparentes	195
5.55	Secuencia 'eβe ed.e eɣe' pronunciada por tres hablantes diferentes	196
5.56	Distribución de frecuencias de cada uno de los formantes tomando los datos referentes a todas las vocales	198
5.57	Distribución de frecuencias del primer formante para las cinco vocales estudiadas	199
5.58	Distribución de frecuencias del segundo formante para las cinco vocales estudiadas	200
5.59	Distribución de frecuencias del tercer formante para las cinco vocales estudiadas	201
5.60	Resultados gráficos del análisis discriminante para vocales	204
5.61	Resultados gráficos del análisis discriminante para hablantes	205
5.62	Resultados obtenidos de la evolución de los formantes alrededor de los sonidos básicos pertenecientes a los sonidos bilabiales [p], [β], [m]	209
5.63	Resultados obtenidos de la evolución de los formantes alrededor de los sonidos básicos pertenecientes a los sonidos dentales [t], [d.], [n]	210
5.64	Resultados obtenidos de la evolución de los formantes alrededor de los sonidos básicos pertenecientes a los sonidos palatales/velares [k], [ɣ], [ŋ]	211
5.65	Resultados obtenidos de la evolución de los formantes alrededor de los sonidos básicos pertenecientes a los sonidos [f], [θ], [s]	212
5.66	Resultados obtenidos de la evolución de los formantes alrededor de los sonidos básicos pertenecientes a los sonidos [x], [tʃ]	213
5.67	Resultados obtenidos de la evolución de los formantes alrededor de los sonidos básicos pertenecientes a los sonidos [l], [λ], [r], [rr]	214

1 INTRODUCCIÓN

El campo de investigación abierto en el área del tratamiento automático de la voz, es amplio y complejo. Es amplio debido a los diversos objetivos que comprende (síntesis de voz, reconocimiento del habla, reconocimiento del hablante, detección de sonidos mal pronunciados en una lengua, etc.), y es complejo porque se intenta replicar simultáneamente la funcionalidad del oído humano y de la parte del cerebro dedicada a descodificar señales auditivas y convertirlas en abstracciones de más alto nivel.

Tratar de imitar al oído y parte del cerebro humano resulta de por sí un objetivo extremadamente ambicioso, pero si a esta circunstancia se le añade la constatación de la gran diversidad que existe en la producción del habla, nos encontramos no sólo ante un difícil objetivo, sino más bien ante un verdadero reto.

Como dificultad añadida, éste es un problema cuya solución requiere la aplicación de conocimientos en diversos campos de la ciencia como son la fonética, tratamiento de señal, informática, matemáticas, neurología, etc.

La naturaleza multidisciplinar de los conocimientos necesarios para abordar soluciones a los diversos objetivos que plantea el tratamiento automático de la voz, complica la creación de equipos reducidos de investigación en el campo. La concepción de esta tesis doctoral está inspirada en el problema mencionado.

La introducción en los últimos años de sofisticados equipos informáticos en los centros de investigación y desarrollo, así como la posibilidad de acceso a equipos similares en los sectores económicos de servicios, industria y en el público en general, ha posibilitado la creación de diversas aplicaciones prácticas que cubren los objetivos principales abiertos en el campo del tratamiento de la voz, sin embargo, ninguna de las aplicaciones conocidas

consigue una aproximación brillante a las posibilidades del ser humano que intentan ser emuladas.

En síntesis de voz, mucho más asequible que el reconocimiento, se han creado programas informáticos capaces de leer texto de una forma comprensible, pero fácilmente detectable como automática. En reconocimiento, IBM ha conseguido una herramienta bastante fiable aplicada a un conjunto finito aunque amplio de palabras pronunciadas aisladamente. En el campo del reconocimiento del hablante no se han conseguido resultados lo suficientemente robustos como para desplazar a los sistemas de autentificación tradicionales, y así sucesivamente en las demás áreas.

Bajando el nivel de abstracción e introduciéndonos en las distintas disciplinas de la ciencia cuyos conocimientos son básicos en las áreas escogidas, nos encontramos con grandes ayudas en la mayoría de ellas:

- El reconocimiento automático paramétrico mediante redes neuronales ha sido ampliamente estudiado para su aplicación en el campo de la voz, aunque con resultados parciales [FRE93], [HIL95b], [NAG94].

- En matemáticas, las cadenas de Markov nos brindan unas características de robustez y seguridad muy apropiadas para su utilización en desarrollo de aplicaciones de reconocimiento de voz [RAB89], [RAB93].

- La utilización de métodos clásicos de tratamiento de señal como la transformada de Fourier, ha sido complementada con otros alternativos como el de predicción lineal (LPC) cuyas características resultan más aconsejables para la consecución de diversos objetivos como por ejemplo disminuir el tiempo de respuesta en aplicaciones de reconocimiento de voz o mejorar la fiabilidad en la detección de formantes [ROW92], [PAR86], [RAB78], [CAN74], [SCH70].

- En el campo de la fonética, se da la circunstancia de que aunque se cumplen principios generales, su estudio es muy dependiente de la lengua concreta cuyas características acústicas se desean obtener, por ello, los avances que se realizan en fonética acústica inglesa o alemana, por ejemplo, no siempre se asimilan de forma directa por la castellana, andaluza, o cualquier otra [QUI93], [MAR94].

Debido a las restricciones expresadas en la asimilación e intercambio de conocimientos en el campo de la fonética acústica, junto a la existencia de un mayor desfase entre los temas propios de los lingüistas por una parte, y los ingenieros, matemáticos, etc. por la otra, se constata que el nivel que existe actualmente en el campo de la fonética acústica española resulta insuficiente para complementar al resto de las disciplinas involucradas en el área del tratamiento automático de la voz [SAB84].

Las publicaciones clásicas en el campo de la fonética acústica española son el fruto de investigaciones realizadas con medios que hoy en día resultan claramente obsoletos [KOE46], [MED85]. El enfoque de estos trabajos tampoco es el idóneo para su aplicación en el campo del tratamiento automático del habla, esto es debido por una parte a su orientación lingüística, y por otra a la carencia de resultados lo suficientemente claros, precisos y generales como para poder ser utilizados en aplicaciones de voz.

Con la realización de esta tesis se pretende minimizar en la medida de lo posible las carencias que existen en el campo de la fonética acústica en cuanto a métodos y herramientas para el análisis de la voz se refiere. Por otra parte, se desea ofrecer documentación detallada de los resultados obtenidos aplicando los algoritmos ideados sobre la fonética castellana.

Las fases fundamentales de las que ha constado la elaboración de la tesis son las que se detallan a continuación:

- En primer lugar se ha realizado un estudio del estado del arte en las áreas principales que deben ser tratadas para cubrir los objetivos propuestos para la tesis.

- El segundo paso consistió en el desarrollo de una aplicación informática que implementa el método de predicción lineal para obtener los parámetros LPC, así como la etapa de traspaso a espectros de voz. Previamente se hizo un estudio para elegir el soporte de toma de datos más apropiado para nuestras necesidades, haciendo especial énfasis en la posibilidad de programación a partir del hardware seleccionado [KEW93], [HEI93].

- Seguidamente se pasó a la fase principal de este trabajo, en la que ha primado la componente de investigación necesaria en toda tesis doctoral. Se han ideado, probado y evaluado distintos métodos de tratamiento de señal para la obtención de características espectrales de la voz, centrándose la atención en la determinación de formantes.

- Los desarrollos informáticos se realizaron siguiendo un esquema de prototipado rápido con soporte de entornos visuales sobre lenguajes orientados a objetos [CHA96], [COA93], [GRA91]. Esta elección, técnicamente es la que más se adapta al mecanismo de prueba y error que se ha seguido para la concepción de los métodos de tratamiento de señal que pretende aportar esta tesis.

- Por último, empleando los resultados conseguidos en la fase anterior, se ha realizado un estudio de la fonética acústica castellana, cuyo objetivo principal consiste en determinar la posición y evolución de los formantes.

La estructura del libro es la siguiente:
En primer lugar se detallan los objetivos principales que se pretenden cubrir con la realización de esta tesis doctoral. Después, en el capítulo dedicado al estado del arte, se explican los métodos básicos de traspaso de señales desde el dominio del tiempo al de la frecuencia. También se exponen algunos conceptos básicos de fonética acústica española y se muestran los resultados de las investigaciones más relevantes de fonética relacionadas con los temas que se tratan en la tesis.

El siguiente capítulo se dedica a la explicación de los trabajos desarrollados en el campo de la caracterización espectral de sonidos, centrándose en la correcta determinación de los formantes del habla. Los sonidos sordos y sonoros se tratan separadamente, y a estos últimos se llega mediante un estudio previo de los segmentos vocálicos.

Después del estudio realizado a nivel de tratamiento de la señal de voz, se dedica un capítulo al análisis de la situación y evolución de los formantes en sonidos de la lengua española; para ello, se emplean los métodos, algoritmos y herramientas desarrollados en la tesis y se abordan diversos trabajos de caracterización espectral de sonidos tanto aislados como coarticulados, en uno o varios hablantes. A continuación se detallan las conclusiones obtenidas, las aportaciones originales y las líneas de ampliación de la tesis. Finalmente, se plasman en la bibliografía los artículos, libros y papeles de congresos referenciados en este trabajo. También se incluye un apéndice con los espectros obtenidos de diversos sonidos coarticulados del español.

2 OBJETIVOS

La **hipótesis** en la que se basa el desarrollo de esta tesis, se centra en la suposición de que partiendo del método de predicción lineal, es posible idear algoritmos de tratamiento de señal que permitan obtener una buena estimación de características espectrales significativas de la voz, especialmente en la detección de los formantes que se producen en el habla. Estos algoritmos ayudarían a construir un catálogo analítico de los principales sonidos del español, con el objetivo de complementar los estudios realizados hasta el momento en el campo de la fonética acústica.

La extracción de formantes es un objetivo fundamental del trabajo debido a la importancia que tiene esta información para conseguir una correcta interpretación de los espectros de voz. También se pretende obtener métodos alternativos que caractericen los sonidos sordos.

Crear herramientas software y unidades reutilizables se considera necesario tanto para el desarrollo de un trabajo de carácter empírico en este campo como para la posterior elaboración de los estudios de fonética acústica.

Con el fin de facilitar la comprensión de los conceptos que se expondrán a lo largo del libro, se pretende documentar la tesis con la ayuda de toda clase de gráficas representativas obtenidas a lo largo del desarrollo de este trabajo.

Una vez finalizada la fase de estimación de características espectrales, se desea realizar un estudio de la fonética acústica castellana usando los métodos ideados y las herramientas desarrolladas.

La completa universalidad de los resultados obtenidos en el estudio fonético no es un objetivo de nuestro trabajo, puesto que esta meta sería lo suficientemente ambiciosa como para generar en sí misma una nueva tesis, sin embargo, se pretende ofrecer los principios generales conseguidos tras realizar un estudio detallado con un número reducido de hablantes. También se ofrecerá la experiencia aportada tras la aplicación de una metodología de trabajo que se espera pueda ser extrapolada a futuras investigaciones en el campo de la fonética acústica empleando un número elevado de hablantes y contextos.

3 ESTADO DEL ARTE

3.1 INTRODUCCIÓN

En este capítulo se pretende aportar una revisión de conocimientos básicos y líneas de investigación que sirvan como punto de partida al desarrollo de la tesis. En primer lugar se describirán dos de los métodos más utilizados para descomponer una señal en frecuencias: la transformada de Fourier y el método de predicción lineal.

La comprensión y utilización de los algoritmos de traspaso de una señal en el dominio del tiempo al dominio de la frecuencia, resulta fundamental para obtener los parámetros básicos a partir de los cuales se podrán realizar caracterizaciones espectrales de los sonidos, reconocimiento del habla, etc.

En el tercer apartado se realiza una descripción de los sonidos básicos del español desde un punto de vista fonético. Para comenzar se presenta un cuadro de sonidos clasificados por el punto y el modo de articulación, después se va revisando cada grupo, explicándose las características articulatorias y acústicas principales.

En el desarrollo de la tesis no se realizará un estudio de todos los sonidos del español, sino que se seleccionarán aquellos que se consideren más representativos para el desarrollo de los trabajos de fonética que se pretenden abordar.

Por último, se realiza una recopilación de ideas y enfoques pertenecientes al campo del tratamiento de la voz desde un punto de vista de la fonética acústica; para ello, se han revisado las publicaciones técnicas que se han considerado más ajustadas a los propósitos iniciales de la tesis. La revista más utilizada ha sido 'Journal of Acoustic Society of America' (JASA), y la conclusión fundamental que se ha obtenido es la importancia que tiene el estudio de la posición y

evolución de los formantes del habla para el desarrollo de la mayor parte de los campos asociados al tratamiento automático de la voz.

En este capítulo habría cabida a revisiones de muchos otros temas relacionados con los estudios que van a ser realizados, sin embargo, se han escogido aquellos que se han considerado fundamentales para el desarrollo de la tesis. Entre los temas que guardan relación con esta tesis, pero que no son tratados de una forma directa por la misma, cabe resaltar los siguientes: fundamentos del tratamiento de la señal [RAB78], [RAB93], [PAR86], [ROW92], [FOL92]. Cuantización vectorial [RAB93], [MAK85]. Cadenas de Markov [RAB89], [RAB93]. Redes neuronales [FRE93], [HIL95], [NAG94]. Representación visual de las señales de voz [COO93]. Síntesis de voz [MAR90b], [ROW92]. Sistemas de representación del conocimiento [HOR85], [JAC90], etc.

3.2. MÉTODOS Y ALGORITMOS BÁSICOS PARA LA OBTENCIÓN DE CARACTERÍSTICAS DE LA SEÑAL DE VOZ

3.2.1 ANÁLISIS EN EL DOMINIO DEL TIEMPO

Aunque para realizar un adecuado análisis de la señal de voz es necesario acudir al dominio de la frecuencia, existen diversas operaciones matemáticas que podemos realizar en el dominio del tiempo que nos proporcionan información importante de las porciones del habla consideradas.

Debido a la propia naturaleza de la voz, resulta conveniente analizar la señal en bloques, de tal forma que podamos establecer características que varíen según lo hace el habla. Para ello, es adecuado establecer un sistema de ventanas que separe y pondere las muestras de la señal de voz que se pretende analizar.

El mecanismo de ventaneo se ilustra en la figura 3.1, a cada porción de la señal de voz (del tamaño deseado) se le asigna una ventana, de tal forma que las muestras queden ponderadas con los valores de la función escogida (en el ejemplo Hamming). En este caso, las muestras que se encuentran en los extremos de la ventana tienen un peso mucho menor que las que se hallan en el medio, lo cual es muy adecuado para evitar que características de los extremos del bloque varíen la interpretación de lo que ocurre en la parte más significativa (central) de las muestras seleccionadas.

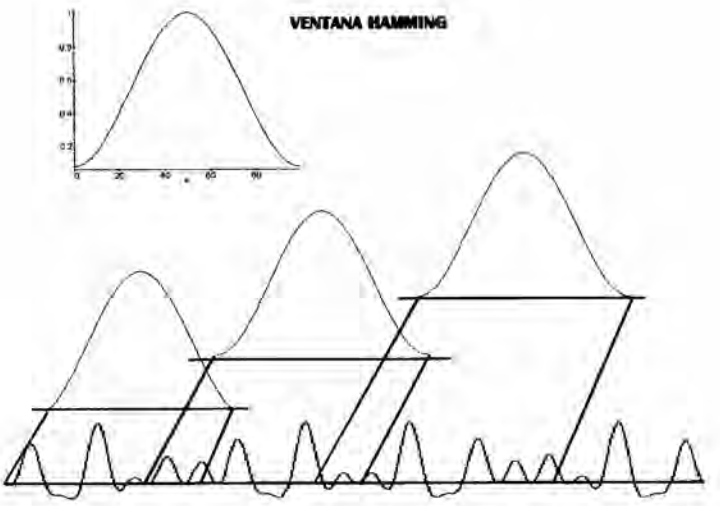


Figura 3.1

Mecanismo de establecimiento de ventanas en la señal a analizar.

Como se puede observar en la figura anterior, la colocación de las ventanas puede realizarse de tal forma que existan solapamientos. Aunque esto repercutirá negativamente en los tiempos de respuesta de los algoritmos utilizados, proporcionará una mejor calidad en los resultados obtenidos.

Las funciones más comunes que se emplean para establecer ventanas son:

- Ventana rectangular:

  $w(t) = 1$ en el intervalo temporal del bloque $(0 \leq t \leq L\text{-}1)$

  $w(t) = 0$ en el resto de la señal

- Ventana Hamming:

  $w(t) = 0.54 - 0.46*\text{coseno}(2\pi t/L)$ en $0 \leq t \leq L\text{-}1$

  $w(t) = 0$ en el resto de la señal

- Ventana Hanning:

  $w(t) = 0.5 - 0.5*\text{coseno}(2\pi t/L)$ en $0 \leq t \leq L\text{-}1$

  $w(t) = 0$ en el resto de la señal

La ventana más utilizada es la Hamming, sin embargo, los valores en sus extremos quedan muy reducidos, por lo que se suele solapar este tipo de ventanas para eliminar el efecto mencionado.

Para cualquier ventana, su duración determina la cantidad de cambios que se podrán obtener, dado que con una duración temporal larga se omiten los cambios locales producidos en la señal, mientras que con una duración demasiado corta se reflejan demasiado los cambios puntuales.

Relacionando el tamaño de las ventanas y la frecuencia de muestreo de la señal, obtenemos un filtrado que dadas las características del aparato fonador habitualmente se sitúa entre los 45Hz. y los 300Hz.

Una vez establecido el mecanismo de ventanas, las características temporales más comunes que se utilizan en la señal de voz son las siguientes:

- **Energía y magnitud**

Tanto la energía como la magnitud son útiles para distinguir segmentos sordos y sonoros en la señal de voz, dado que, los valores de ambas características aumentan en los sonidos sonoros respecto a los sordos.

En la gráfica superior de la figura 3.2, se representa la función de magnitud de la palabra 'hipotenusa'. Como se puede observar, los valores de mayor energía se corresponden con los segmentos vocálicos de la señal, mientras que en las consonantes oclusivas ocurre lo contrario.

Formulación de la magnitud: $$M(n) = \frac{1}{N}\sum_{m=0}^{N-1}|x(m)| * w(n-m)$$

Formulación de la energía: $$E(n) = \frac{1}{N}\sum_{m=0}^{N-1}x(m)^2 w(n-m)$$

Estas funciones nos dan una idea de la amplitud de la señal en un intervalo considerado, por ello su valor aumenta en los sonidos sonoros, en los que el aire encuentra menos impedimentos para salir de los órganos articulatorios.

- **Cruces por cero y máximos**

Los cruces por cero indican el número de veces que una señal continua toma el valor cero. Para las señales discretas, un cruce por cero ocurre cuando dos muestras consecutivas difieren de signo, o bien una muestra toma el valor nulo.

Habitualmente, las señales con mayor frecuencia presentan un mayor valor en esta característica, el ruido también genera un gran número de pasos por cero, por lo que una utilización práctica consiste en analizar las señales grabadas desde esta óptica para comprobar su calidad.

Desde un punto de vista acústico, con los cruces por cero se puede intentar detectar las fricaciones del habla. En la imagen central de la figura 3.2 se representa la gráfica que, sin ningún genero de dudas, localiza la fricación de la 's' en la palabra 'hipotenusa'.

El problema que presentan los cruces por cero es la sensibilidad que se da a las componentes continuas de la señal. Podemos encontrar un estimador alternativo contabilizando los máximos (o mínimos) que existen en la señal de voz. La imagen inferior de la figura 3.2 representa esta característica, que como se puede apreciar, resulta muy adecuada para diferenciar los sonidos sonoros del resto de la señal (incluido el silencio).

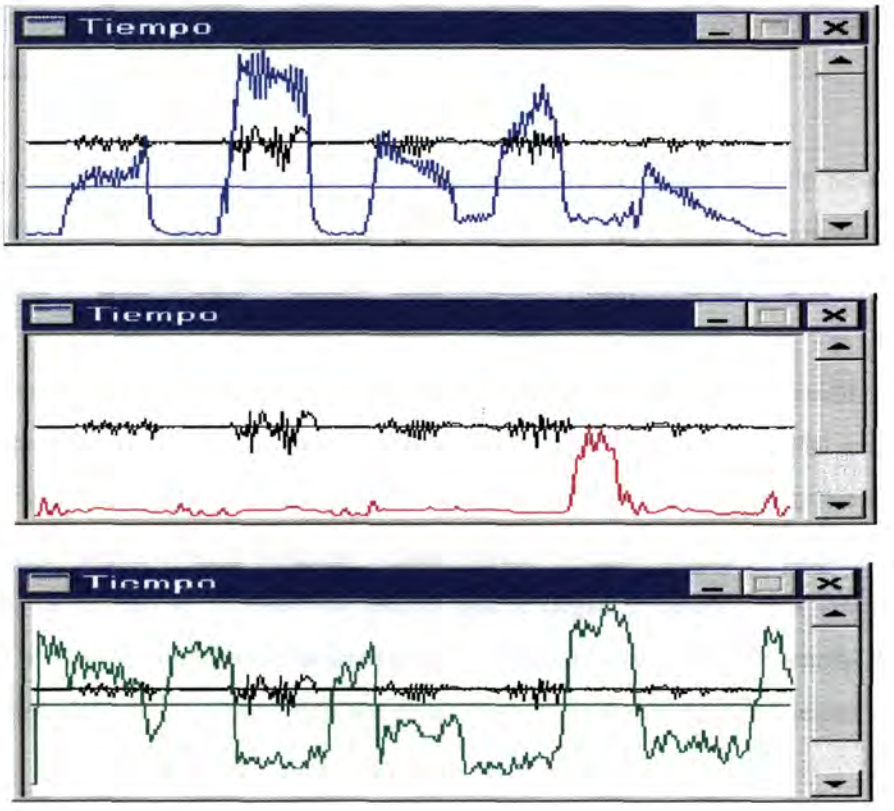


Figura 3.2

Funciones de energía, cruces por cero y máximos en la palabra 'hipotenusa'.

3.2.2 ANÁLISIS EN EL DOMINIO DE LA FRECUENCIA

En este apartado se describen someramente los dos métodos matemáticos más conocidos y utilizados de traspaso de señal al dominio de la frecuencia, estos son la transformación de Fourier [FOL92], [COC67], [RAB78] y el método de predicción lineal (LPC) [MAK75], [CAN74], [ATA71], [PAR86]. En el caso de la transformada de Fourier, se pretende únicamente explicar cómo y por qué funciona, sin entrar en el formalismo matemático que lo sustenta, con ello se espera introducir en el concepto a aquellas personas que leyendo esta tesis no desean profundizar en este aspecto de un campo (el tratamiento de la voz), de carácter marcadamente multidisciplinar.

Otra decisión ha sido enfocar la explicación de la transformada de Fourier desde el formulismo de la señal discreta, más cercano a su utilización en métodos y algoritmos computables, con su vertiente práctica de creación de herramientas y aplicaciones en el campo tratado.

Antes de comenzar describiendo ningún método concreto, resulta adecuado incidir en la idea básica subyacente en estos procesos, esto es, la descomposición de una señal compleja en sumatorio de señales simples. El oído humano, por medio del caracol, descompone las señales auditivas que le llegan en sus frecuencias fundamentales [MAR94], y ésta es la información básica a partir de la cual se elaboran las señales que le llegan al cerebro. Por tanto podemos afirmar que el proceso de audición se fundamenta en la descomposición en frecuencias de la señal sonora. Para entender este concepto mejor nos basaremos en las tres figuras siguientes.

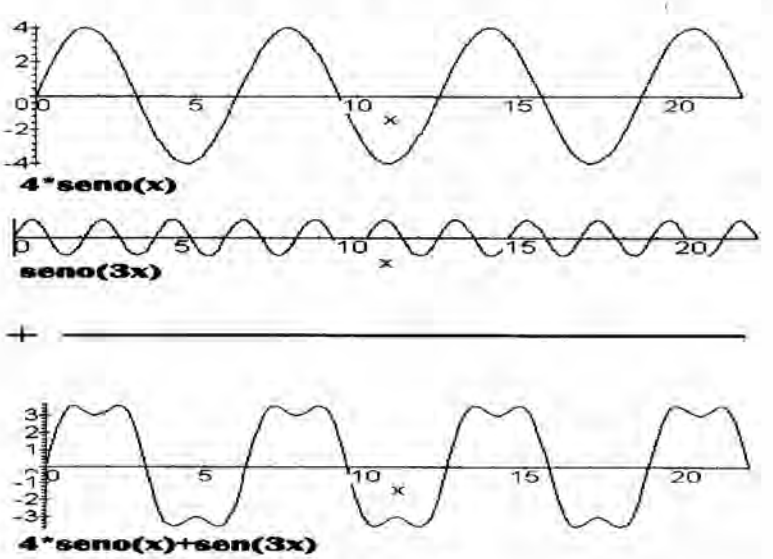


Figura 3.3
Ejemplo de descomposición de una señal compleja en sumatorio de señales simples.

La figura 3.3 nos muestra como la señal compleja representada en la parte inferior se puede descomponer en las dos señales simples de la parte superior, o desde otro punto de vista, la señal inferior puede ser creada sumando las dos funciones sinusoidales superiores. En este ejemplo, se podría reconocer a simple vista las frecuencias y amplitudes que dan lugar a la señal compleja, sin embargo, complicando la señal un poco más, (en la figura 3.4) es difícil determinar la descomposición en senos (o cosenos) de la función final.

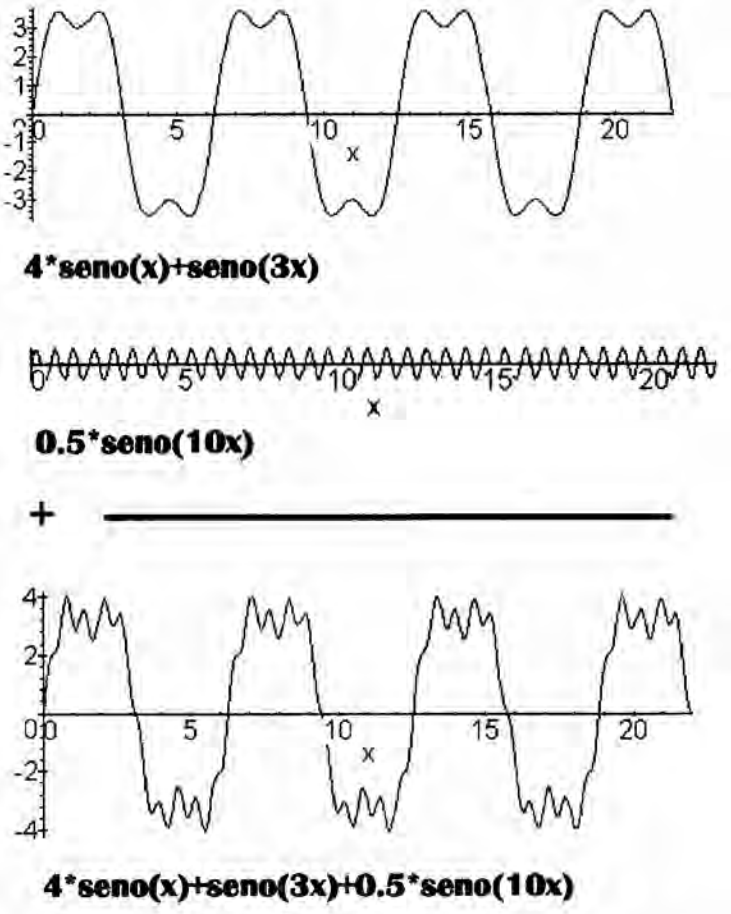


Figura 3.4
Ejemplo de descomposición de una señal compleja en sumatorio de señales simples (continuación).

En la figura 3.5 se presenta una señal periódica que se puede descomponer en 5 señales sinusoidales de diferentes frecuencias y amplitudes. Llegados a este punto conviene remarcar que cuando analizamos una señal de voz en frecuencias, la información que consideramos relevante son las frecuencias de las señales simples con mayor amplitud (mayor peso en la determinación de la señal de voz) que se obtienen de la descomposición de la señal de voz original.

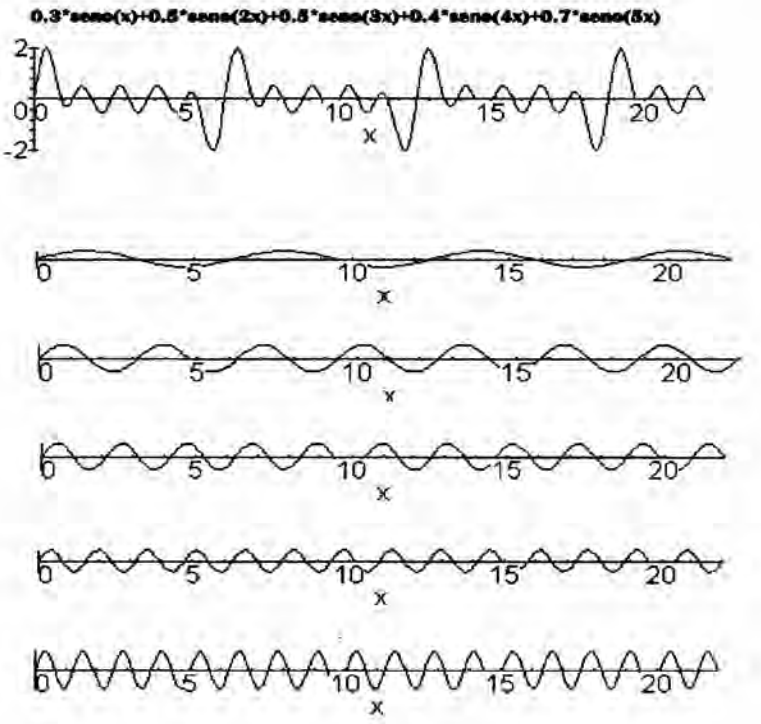


Figura 3.5
Ejemplo de descomposición de una señal compleja en sumatorio de señales simples (continuación).

En el primer método que se explicará (la transformada de Fourier), conviene recordar los conceptos expuestos, puesto que su fundamento estriba en la comparación de diversas ondas sinusoidales y cosenoidales simples con la señal de voz, cuanto más coincida una onda simple con la señal compleja, mayor importancia tiene su frecuencia en la determinación de la señal original.

3.2.2.1 TRANSFORMADA DE FOURIER. CONCEPTOS BÁSICOS

Para explicar el funcionamiento de la transformada de Fourier, partiremos de su formulación básica y explicaremos los conceptos fundamentales.

$$F\left(\frac{n}{NT}\right) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT)\, e^{-j\frac{2\pi nk}{N}} \qquad n = 0,1, \ldots N-1$$

FORMULACIÓN DE LA TRANSFORMADA DE FOURIER

Donde 'N' es el número de muestras de la ventana que se va a analizar, 'T' es el periodo de muestreo (inverso a la frecuencia de muestreo que denominaremos 'f'), 'n' es el índice de la frecuencia cuyo valor queremos obtener y '$m(kT)$' indica la muestra tomada en el instante 'kT' (muestra Késima) de la ventana. Algunos de estos valores se ilustran en la figura 3.6.

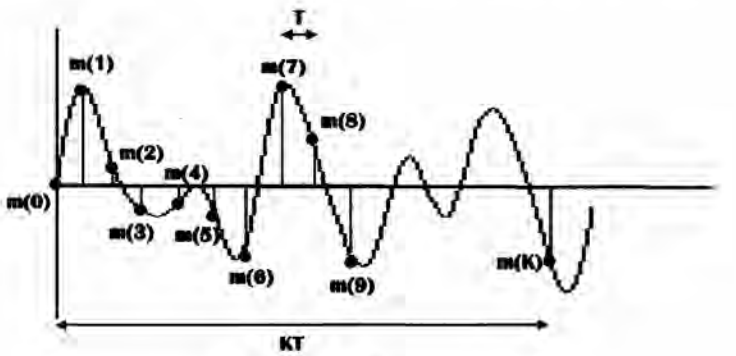


Figura 3.6

Señal muestreada, con indicaciones del significado de 'm(k)', 'T' y 'kT'.

El valor del parámetro 'n' determina la frecuencia concreta que se va a analizar, es decir, representa una de las frecuencias en las que se va a tratar de descomponer la señal de partida, de esta manera, para hacer el estudio con todas las frecuencias, usamos el rango completo de variación de 'n': $n=0,1,2, \ldots N-1$.

Desarrollando la fórmula de la transformada de Fourier para los distintos valores de 'n', tenemos:

$$n = 0 \Rightarrow \quad F(0) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT) * e^{0}$$

$$n = 1 \Rightarrow \quad F(\frac{1}{N} f) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT) * e^{-j2\pi\frac{k}{N}}$$

$$n = 2 \Rightarrow \quad F(\frac{2}{N} f) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT) * e^{-j4\pi\frac{k}{N}}$$

.....

$n = N\text{-}1 \quad \Rightarrow \quad F(\frac{N-1}{N} f) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT) * e^{-j2(N-1)\pi\frac{k}{N}}$

Con las ecuaciones anteriores se obtienen las siguientes ideas:

- La porción de la señal que se analiza se encuentra en el bloque de *muestras m(0), m(1), m(2), ..., m(N-1),* debido al sumatorio $\sum_{k=0}^{N-1} m(kT)*\cdots$

- La frecuencia *f=0Hz.* (correspondiente a *n=0*) se halla haciendo la media aritmética de los valores de las muestras, por lo que representa la componente continua de la señal.

- El parámetro *'n'* actúa de índice para obtener las distintas frecuencias de estudio, por ello nos encontramos con la secuencia: $F\left(\frac{0}{N}f\right), F\left(\frac{1}{N}f\right), F\left(\frac{2}{N}f\right), \ldots\ldots, F\left(\frac{N-1}{N}f\right)$, donde *n/N* es una proporción (en este caso de *'f'*).

- Los valores que se obtienen para $0 \leq n < N/2$ coinciden con los obtenidos en el intervalo $N/2 \leq n < N\text{-}1$, (con *'n'* par) por lo que es suficiente realizar los cálculos en una de las dos mitades. Según el .criterio de Nyquist, el ancho de banda de la señal coincide con la mitad de la frecuencia de muestreo *'f'*, correspondiente a *n=0, 1, 2, ..., (N/2)-1.*

- Los valores $n=1 \Rightarrow 2\pi$, $n=2 \Rightarrow 4\pi$, ..., $n=N\text{-}1 \Rightarrow 2(N\text{-}1)\pi$, indican las frecuencias de las señales sinusoidales y cosenoidales con las que se comparará la señal original, este concepto se explicará a lo largo del apartado.

- Si aumentamos el valor de *'N'*, conseguimos hacer el análisis con un mayor número de frecuencias $(0 \leq n \leq N\text{-}1)$, pero a costa de un mayor tiempo para calcular las operaciones del sumatorio $(0 \leq k \leq N\text{-}1)$.

A modo de ejemplo, realizaremos el mismo desarrollo suponiendo una señal de voz muestreada a 10000 muestras/sg. y un bloque de 100 datos *(N=100)*. Esto implica que se va a realizar el análisis de 10 ms. de tiempo.

$$n = 0 \Rightarrow \quad F(\frac{0}{100}10000) = F(0Hz.) = \frac{1}{100}\sum_{k=0}^{99} m(kT)*e^{0}$$

$$n = 1 \Rightarrow \quad F(\frac{1}{100}10000) = F(100Hz.) = \frac{1}{100}\sum_{k=0}^{99} m(kT)*e^{-j2\pi\frac{k}{100}}$$

$$n = 2 \Rightarrow \quad F(\frac{2}{100}10000) = F(200Hz.) = \frac{1}{100}\sum_{k=0}^{99} m(kT)*e^{-j4\pi\frac{k}{100}}$$

.....

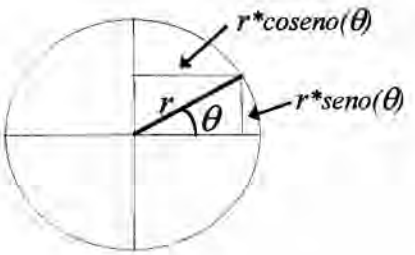
$$n = 49 \Rightarrow \quad F(\frac{49}{100}10000) = f(4900Hz.) = \frac{1}{100}\sum_{k=0}^{99} m(kT)*e^{-j98\pi\frac{k}{100}}$$

Como se puede observar, el análisis se realiza sobre un ancho de banda de 5KHz. con una resolución espectral de 100 Hz.

Una vez explicados los conceptos básicos relativos a anchos de banda, resolución espectral, etc. del análisis, vamos a pasar a detallar el significado y funcionamiento del sumatorio de la fórmula de la transformada de Fourier.

Con un *'n'* (frecuencia) y *'N'* fijados, el sumatorio: $\sum_{k=0}^{N-1} m(kT)*e^{-j2\pi n\frac{k}{N}}$ depende únicamente del parámetro *'k'*. En este sumatorio la operación primordial es la multiplicación de cada muestra *m(kT)* del bloque de datos por el valor exponencial. Esta multiplicación, como veremos más tarde, hace las funciones de comparador entre ambos operandos (*m(kT)* y el valor exponencial).

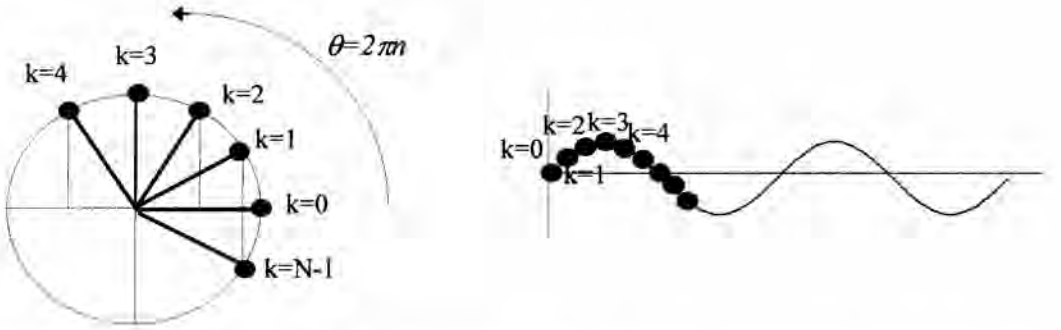
La expresión exponencial representa un número complejo en coordenadas polares $(r*e^{j\theta})$, donde *'r'* es el módulo y *'θ'* el argumento. Gráficamente:



$$r*e^{j\theta} = r(\cos\theta + j*sen\theta)$$

Por lo que: $e^{-j2\pi n\frac{k}{N}} = \cos\left(-2\pi n\frac{k}{N}\right) + j*\text{sen}\left(-2\pi n\frac{k}{N}\right)$

Puesto que el argumento del valor complejo de nuestra expresión depende de *'k'*, esto significa que nos encontramos ante un complejo que va girando en el círculo unidad con una velocidad angular $\theta=2\pi n$ y con saltos discretos de k/N radianes del recorrido total. La representación gráfica de este concepto se expresa a continuación:



La rotación del valor complejo en el círculo unidad se puede representar mediante sus correspondientes valores de senos y cosenos, de manera que el sumatorio:

$$\sum_{k=0}^{N-1} m(kT) * \left\{ \cos\left(-2\pi n \frac{k}{N}\right) + j * \text{sen}\left(-2\pi n \frac{k}{N}\right) \right\}$$

se convierte en un recorrido en *'k'* de cada muestra del bloque por cada valor de la función seno (y coseno) correspondiente *{m(0)*sen(0)+m(1)*sen(-2πn/N)+m(2)*sen(-4πn/N)+ ... }* y análogamente con los cosenos.

Esta idea se representa en la figura 3.7, en la que aparecen tres senos con frecuencias 2π *(n=1)*, 4π (n=2) y 6π (n=3). Estos senos se comparan con un bloque (de tamaño *'N'*) de la señal que se pretende descomponer. A simple vista se adivina que el seno que mejor encaja es el segundo*(4π)*.

La razón por la que el sumatorio actúa de comparador (y este razonamiento es clave para entender el funcionamiento de la transformada de Fourier), es que cuando el bloque de la señal analizada es parecido al seno (o coseno) por el que se multiplican las muestras, el valor final del sumatorio será alto (alejado de cero), esto es así porque los valores positivos del seno (o coseno) se multiplicarán por valores positivos de la señal, y los valores negativos por los negativos de la señal. En el caso de que la señal no se parezca al seno (o coseno), los valores positivos y negativos se irán contrarrestando, y el resultado del sumatorio se aproximará más a cero.

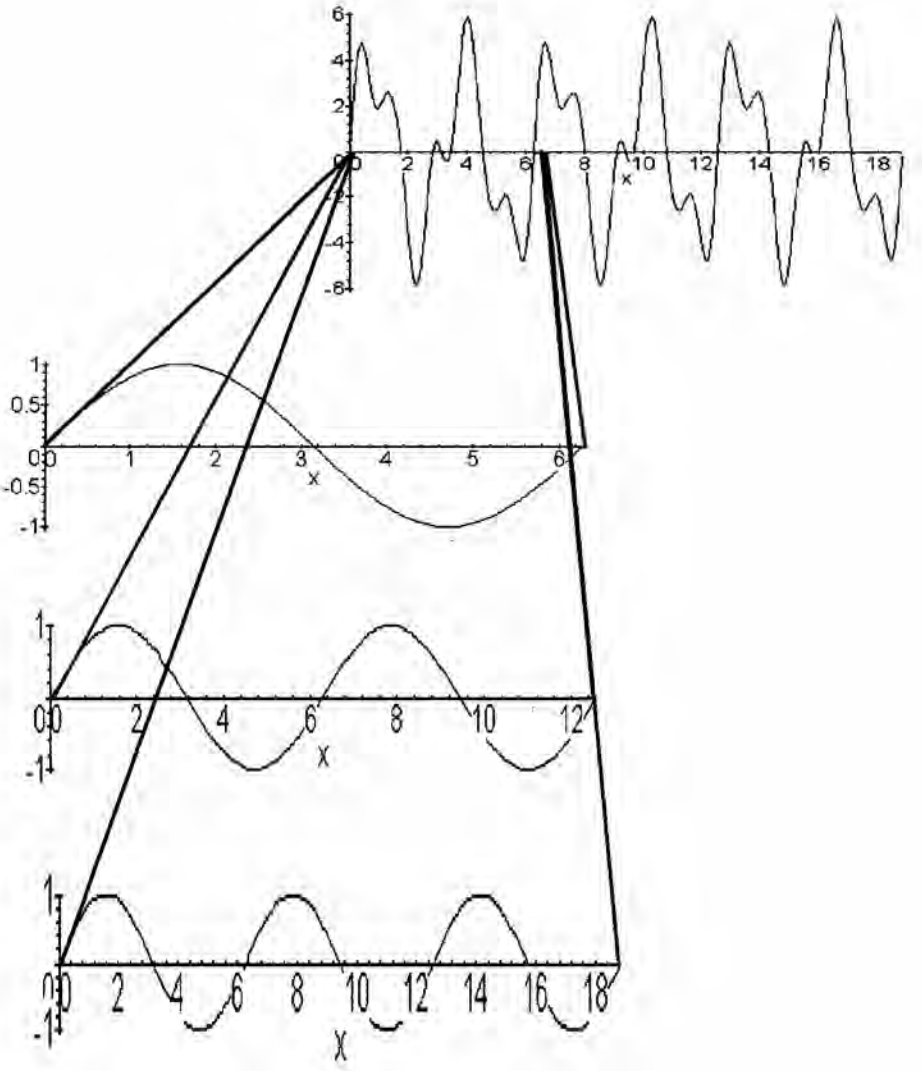


Figura 3.7

Comparación de funciones seno de diferentes frecuencias con un bloque de la señal que se pretende descomponer.

Las figuras 3.8 y 3.9 muestran el resultado de comparar una señal compleja (función superior de la figura 3.8) por señales sinusoidales de diferentes frecuencias. Como se puede apreciar, cuando la frecuencia del seno forma parte de la composición de la señal original (en nuestro caso *seno(x)*, *seno(3x)* y *seno(10x)*), el resultado del sumatorio (área de la función) se hace bastante mayor que cero, mientras que si la frecuencia del seno no forma parte de la señal original (el resto de los casos), el valor del sumatorio se aproxima a cero.

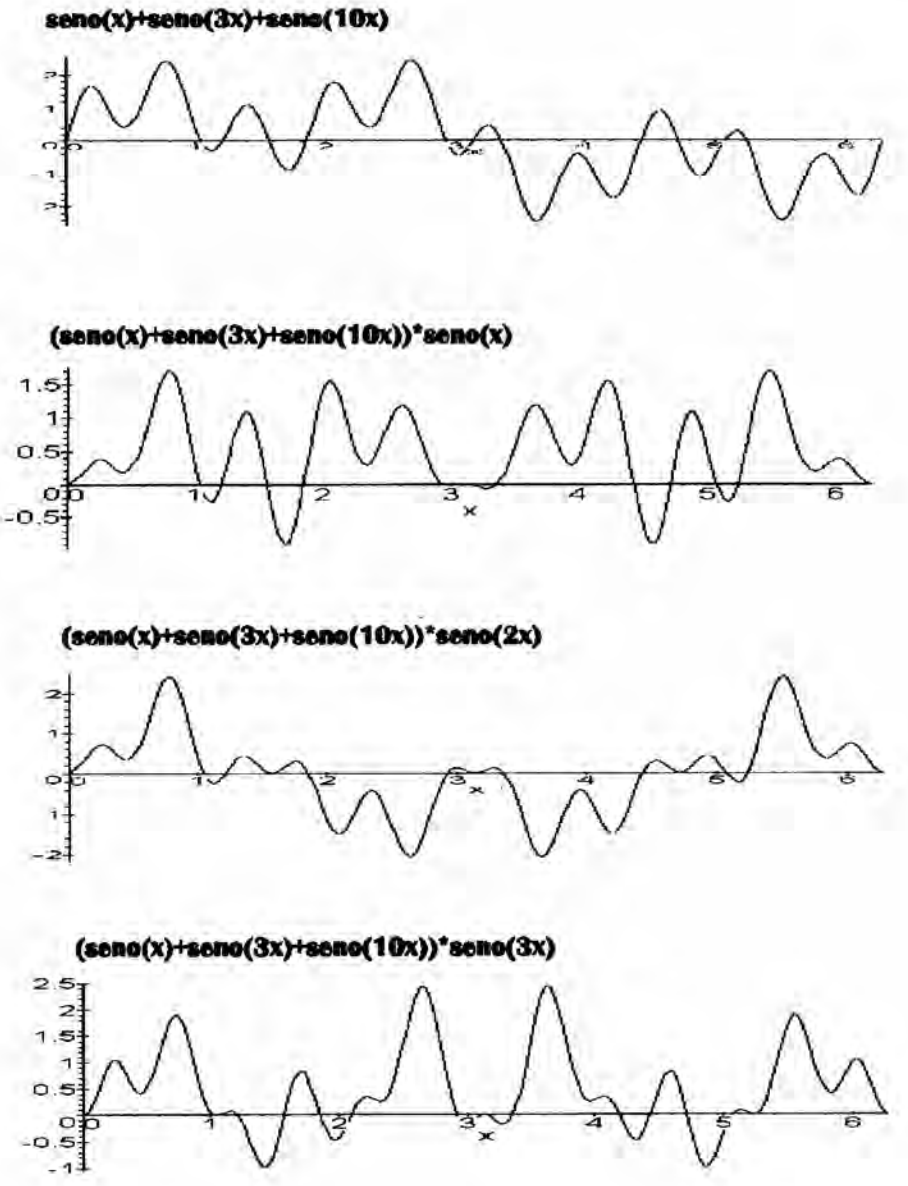


Figura 3.8

Señal que se pretende descomponer y resultado de compararla con senos de diferentes frecuencias.

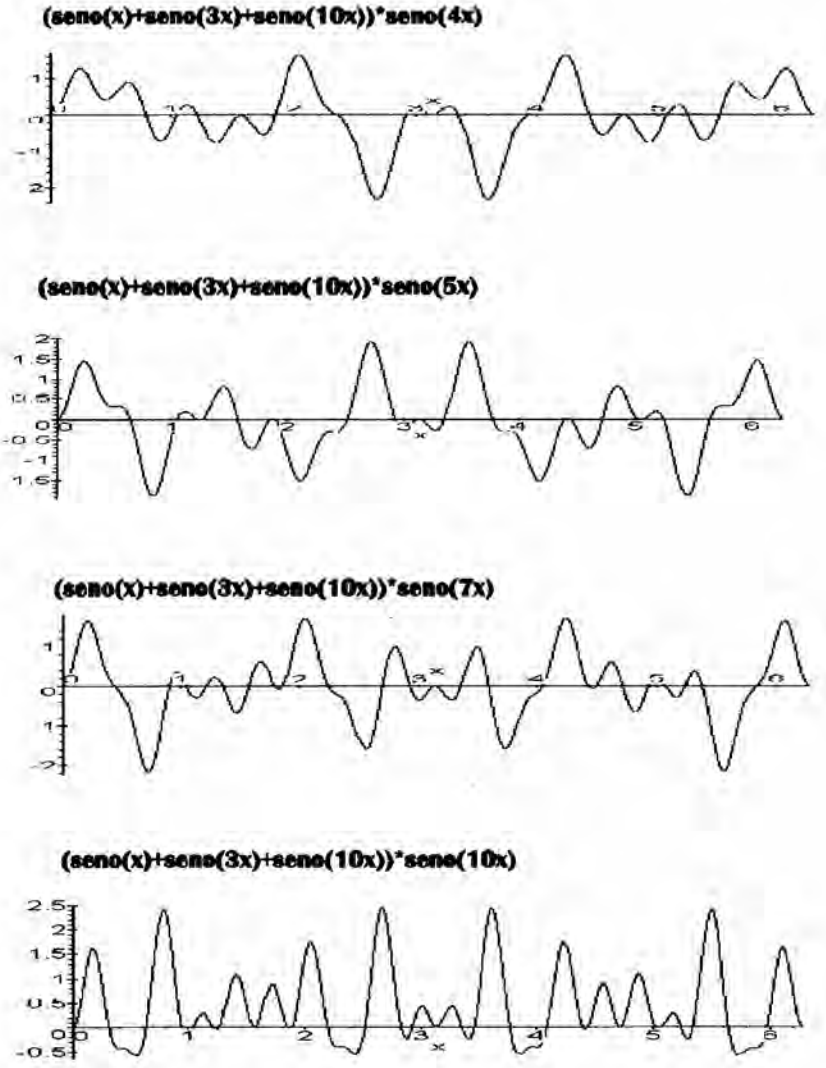
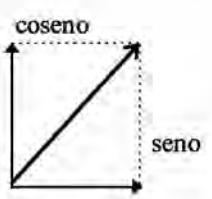


Figura 3.9

Señal que se pretende descomponer y resultado de compararla con senos de diferentes frecuencias (continuación).

El resultado del sumatorio (para un *'n'* fijado) es un número complejo que indica la similitud de la señal analizada con el seno y coseno de la frecuencia dependiente de *'n'*. La parte real del complejo representa la semejanza entre la señal y el coseno, la parte imaginaria se refiere al seno.



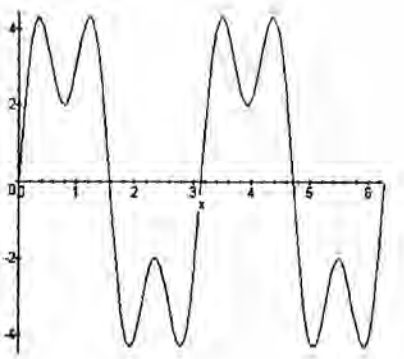
Obviamente un valor alto en el seno o en el coseno indica que la señal original se puede descomponer en un conjunto de señales simples entre las que se encontrará una de esta frecuencia. Para ponderar la importancia de ambos valores, habitualmente se utiliza la distancia euclídea:

$$Modulo = \sqrt{Real^2 + Imag^2}$$

Como ejemplo de los conceptos explicados, vamos a realizar la descomposición en frecuencias de la señal discreta definida por las muestras:

m(0) = 0	m(1) = 4.25	m(2) = 2.63	m(3) = 2.62	m(4) = 4.25
m(5) = 0	m(6) = -4.25	m(7) = -2.63	m(8) = -2.62	m(9) = -4.25
m(10)=0	m(11)= 4.25	m(12)= 2.63	m(13) = 2.62	m(14) = 4.25
m(15) = 0	m(16) = -4.25	m(17) = -2.63	etc.	

- Estas muestras han sido tomadas de la función: *4*sen(2x) + 2*sen(6x).*
- La frecuencia de muestreo *'f'* escogida es 20Hz. *(T=1/20sg.).*
- Puesto que *f=20Hz.* el máximo ancho de banda que se puede analizar es de *10Hz.* (Teorema de Nyquist). La frecuencia máxima de la señal original es de *6Hz.*



Aplicando la transformada de Fourier con *N=10*, y suponiendo que usamos una ventana rectangular que comienza en la muestra *m(7)*:

En primer lugar veremos las posibles frecuencias de estudio en función del parámetro 'n':

$$n=0 \Rightarrow F\left(\frac{0}{10}20\right) = F(0Hz.)$$

$$n=1 \Rightarrow F\left(\frac{1}{10}20\right) = F(2Hz.)$$ (Frecuencia componente de la onda original)

$$n=2 \Rightarrow F\left(\frac{2}{10}20\right) = F(4Hz.)$$

$$n=3 \Rightarrow F\left(\frac{3}{10}20\right) = F(6Hz.)$$ (Frecuencia componente de la onda original)

$n=4 \Rightarrow F\left(\frac{4}{10}20\right) = F(8Hz.)$

[n=0..(N/2)-1]

Vamos a realizar (por ejemplo) el desarrollo de la frecuencia *2Hz. (n=1)*, en donde esperamos encontrar un valor que indique que esta frecuencia es importante en la composición de la señal original.

n=1

$$F(2Hz.) = m(7)\left[\cos\left(-2\pi 1\frac{0}{10}\right) + j\,sen\left(-2\pi 1\frac{0}{10}\right)\right] +$$

$$m(8)\left[\cos\left(-2\pi 1\frac{1}{10}\right) + j\,sen\left(-2\pi 1\frac{1}{10}\right)\right] +$$

.....

$$m(16)\left[\cos\left(-2\pi 1\frac{9}{10}\right) + j\,sen\left(-2\pi 1\frac{9}{10}\right)\right]$$

$$F(2Hz.) = -2.63\,[\cos(0)+j\,sen(0)] + (-2.62)[\cos(-0.628)+j\,sen(-0.628)] +$$
$$-4.25\,[\cos(-1.256)+j\,sen(-1.256)] + (0)[\cos(-1.884)+j\,sen(-1.884)] +$$
$$4.25\,[\cos(-2.512)+j\,sen(-2.512)] + 2.63[\cos(-3.14)+j\,sen(-3.14)] +$$
$$2.62\,[\cos(-3.768)+j\,sen(-3.768)] + 4.25[\cos(-4.396)+j\,sen(-4.396)] +$$
$$0\,[\cos(-5.024)+j\,sen(-5.024)] + (-4.25)[\cos(-5.652)+j\,sen(-5.652)]$$

$$F(2Hz.) = (-18.996+j6.231) = \sqrt{(-18.996)^2 + (6.231)^2} = \boxed{20}$$

El valor obtenido es 20 (si lo ponderamos con el *1/N* de la fórmula nos quedaría 2).

Ahora realizaremos el desarrollo para *n=2 (4Hz.)*, donde esperamos un valor cercano a cero que nos indique que esa frecuencia no es constitutiva de la señal original.

n=2

$$F(4Hz.) = m(7)\left[\cos\left(-2\pi 2\frac{0}{10}\right) + j\,sen\left(-2\pi 2\frac{0}{10}\right)\right] +$$

$$m(8)\left[\cos\left(-2\pi 2\frac{1}{10}\right) + j\,sen\left(-2\pi 2\frac{1}{10}\right)\right] +$$

.....

$$m(16)\left[\cos\left(-2\pi 2\frac{9}{10}\right) + j\,sen\left(-2\pi 2\frac{9}{10}\right)\right]$$

$$F(4Hz.) = -2.63\,[\cos(0)+j\,sen(0)] + (-2.62)[\cos(-1.256)+j\,sen(-1.256)] +$$
$$-4.25\,[\cos(-2.512)+j\,sen(-2.512)] + (0)[\cos(-3.768)+j\,sen(-3.768)] +$$
$$4.25\,[\cos(-5.029)+j\,sen(-5.029)] + 2.63[\cos(-6.28)+j\,sen(-6.28)] +$$
$$2.62\,[\cos(-7.536)+j\,sen(-7.536)] + 4.25[\cos(-8.792)+j\,sen(-8.792)] +$$
$$0\,[\cos(-10.048)+j\,sen(-10.08)] + (-4.25)[\cos(-11.304)+j\,sen(-11.304)]$$

$$F(4Hz.) = (0.018+j0.011) = \sqrt{(0.018)^2+(0.011)^2} = 0.011$$

El valor obtenido es muy cercano a cero. Esto significa que al descomponer la señal original en sus frecuencias básicas, 4Hz. no aparece.

Realizando el mismo desarrollo para las demás frecuencias obtenemos:

$n = 0 \rightarrow 0.020$, $n = 3 \rightarrow 10$, $n = 4 \rightarrow 0.025$

Como se observa, las frecuencias integrantes de la señal original son:

- *n = 1 (f=2Hz.)* → Importancia 20 *(4*sen(2x))*
- *n = 3 (f=6Hz.)* → Importancia 10 *(2*sen(6x))*

Las demás frecuencias se pueden considerar inexistentes en la señal original.

Ahora vamos a modificar la señal de estudio para comprobar el funcionamiento de la transformada de Fourier en diversos casos básicos.

Función: *2*sen(2x) + 8*sen(6x)*

Resultados:

n = 0 (0Hz.) → 0.023
*n = 1 (2Hz.) → 9.983 [2*sen(2x)]*
n = 2 (4Hz.) → 0.006
*n = 3 (6Hz.) → 40 [8*sen(6x)]*
n = 4 (8Hz.) → 0.023

Lo que muestra como se refleja la importancia de cada onda componente de la señal original.

Función: *4*sen(2x) + 2*cos(6x)*

Resultados:

n = 0 (0Hz.) → 0.005
*n = 1 (2Hz.) → 20 [4*sen(2x)]*
n = 2 (4Hz.) → 0.003
*n = 3 (6Hz.) → 10 [2*sen(6x)]*
n = 4 (8Hz.) → 0.004

El método también funciona con funciones más complejas (por ejemplo formadas con senos y cosenos).

Función: *4*sen(2x) + 2*sen(6.5x)*

Resultados:

n = 0 (0Hz.) → 1.66
*n = 1 (2Hz.) → 18.554 [4*sen(2x)]*
n = 2 (4Hz.) → 2.557
*n = 3 (6Hz.) → 9.778 [2*sen(6.5x)]*
n = 4 (8Hz.) → 2.115

En este caso, al no coincidir la frecuencia base de *6.5Hz.* con los valores buscados para cada *'n'*, existe un fenómeno de traspaso de los resultados hacia las frecuencias adyacentes a *6.5Hz.*

Función: *4*sen(2x) + 2*sen(7x)*

Resultados:

n = 0 (0Hz.) → 0.306
*n = 1 (2Hz.) → 20.513 [4*sen(2x)]*
n = 2 (4Hz.) → 1.260
*n = 3 (6Hz.) → 5.327 [2*sen(7x)]*
n = 4 (8Hz.) → 7.759

Al igual que en el caso anterior, sería necesario más definición espectral para detectar la frecuencia de *7Hz.* entre la de *6Hz. (n=3)* y 8Hz. *(n=4)*.

Función: *4*sen(2x) + 2*sen(6x) + 6*sen(8x)*

Resultados:

n = 0 (0Hz.) → 0
*n = 1 (2Hz.) → 19.97 [4*sen(2x)]*
n = 2 (4Hz.) → 0.036
*n = 3 (6Hz.) → 9.969 [2*sen(6x)]*
*n = 4 (8Hz.) → 30 [6*sen(8x)]*

Nueva comprobación del funcionamiento de la transformada de Fourier.

Una vez estudiado el funcionamiento de la transformación de Fourier y realizado un ejemplo numérico, resulta muy sencillo comprender una plantilla de código que implemente la funcionalidad buscada. Las siguientes líneas de código de alto nivel, realizan la transformada de Fourier de una serie de *'N'* muestras *'m(k)'*, con *'N'* par.

```
pi = 3.14
BUCLE n = 0 TO (N/2)-1
COMIENZO
  Sumac ← 0
  Sumas ← 0
  BUCLE k = 0 TO N-1
    Sumac ← Sumac + m(k) * COS(-2 * pi * n * k / N)
    Sumas ← Sumas + m(k) * SIN(-2 * pi * n * k / N)
  FIN (* k (bucle del sumatorio) *)
  Modulo[n] ← SQR(Sumac * Sumac + Sumas * Sumas)
FIN (* n (frecuencias) *)
```

Código para realizar la transformada de Fourier

El resultado queda almacenado en la matriz Modulo[N], sus valores habitualmente son aplanados por una función logaritmo decimal, para evitar que el rango de soluciones sea muy amplio.

Resulta necesario aclarar que el coste computacional requerido para realizar la transformada de Fourier tal y como aquí se ha descrito es muy alto. Cuando se programan implementaciones sobre computadoras, se emplea habitualmente la transformada rápida de Fourier (FFT). Este método se basa en la minimización de los cómputos realizados, evitándose la repetición de los cálculos que se emplean en las distintas etapas de la transformada tradicional.

Una vez explicados los conceptos fundamentales del funcionamiento de la transformada de Fourier, se pasará a desarrollar la teoría que nos dirige a la expresión matemática tomada como punto de partida en este apartado.

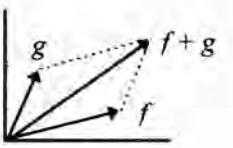
3.2.2.2 TRANSFORMADA DE FOURIER. FORMALISMO

Una vez aclarados los conceptos básicos y el funcionamiento de la transformada de Fourier, vamos a realizar la formalización matemática que da origen a la ecuación de partida del apartado anterior. Para ello nos basaremos en los espacios de Hilbert (espacios infinito-dimensionales), que nos brindan un camino simple y elegante para obtener la expresión básica de traspaso de señales en el dominio del tiempo al dominio de la frecuencia mediante la transformación de Fourier.

Para comenzar, repasaremos algunas propiedades de los espacios n-dimensionales que nos servirán como referencia para entender y emplear sus equivalentes en los espacios de Hilbert. Cada propiedad se ilustra con un gráfico bidimensional

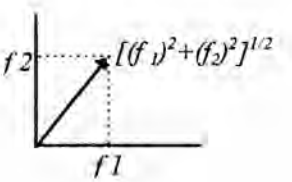
Sean $f = [f_1, f_2, \ldots, f_n]$ y $g = [g_1, g_2, \ldots, g_n]$ dos vectores n-dimensionales:

$f + g = [f_1+g_1, f_2+g_2, \ldots, f_n+g_n]$ *(1)*



$\lambda f = [\lambda f_1, \lambda f_2, \ldots, \lambda f_n]$, $\lambda \in \Re$ *(2)*

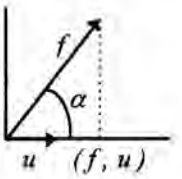
Longitud de f $= \sqrt{f_1^2 + f_2^2 + \cdots + f_n^2}$ (3)



Producto escalar: $(f, g) = f_1 * g_1 + f_2 * g_2 + \ldots + f_n * g_n$ *(4)*

Producto escalar: $(f, g) = \sqrt{f_1^2 + f_2^2 + \cdots + f_n^2}\ \sqrt{g_1^2 + g_2^2 + \cdots + g_n^2}\ \cos(\alpha)$ *(5)*

En el caso de que un vector *'u'* tenga como longitud la unidad, se puede observar con claridad como el producto escalar de un vector *'f'* por *'u'*: (f, u) se corresponde con la proyección de *'f'* sobre el eje determinado por *'u'*.



Como caso particular de suma importancia se encuentra la condición de ortogonalidad (perpendicularidad) entre vectores. Dados dos vectores no nulos *'f'* y *'g'*, se dice que son ortogonales si su producto escalar es nulo: $(f, g) = 0 \Rightarrow cos(\alpha)=0 \Rightarrow \alpha=90^{o}$.

Partiendo de (4) se pueden establecer las siguientes propiedades del producto escalar:

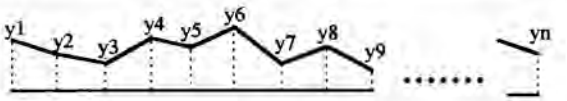
$$(f, g) = (g, f) \qquad (6)$$

$$(\lambda f, g) = \lambda (f, g) \qquad (7)$$

$$(f, g_1 + g_2) = (f, g_1) + (f, g_2) \qquad (8)$$

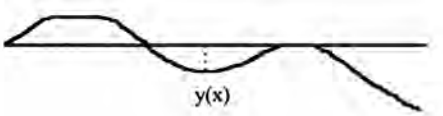
$$(f, f) \geq 0 \qquad (9)$$

Podemos utilizar las características de los espacios n-dimensionales para trabajar con funciones discretas, para ello basta con considerar las distintas muestras de la función como componentes de un vector multidimensional, tal y como se representa en el siguiente gráfico:



$f = [y_1, y_2, \ldots, y_n]$

Los espacios de Hilbert son infinito-dimensionales, por lo que nos encontramos con valores de *'n'* tendiendo a infinito y distancia entre componentes tendiendo a cero. En este caso estamos preparados para trabajar con funciones continuas, de hecho, un vector en un espacio de Hilbert se define como una función $f(x) \ / \ x \in [a,b]$.



En espacios infinito-dimensionales, la suma de vectores y la multiplicación de un vector por un número se definen como la adición de funciones y la multiplicación de una función por un número.

En (3) se define la longitud de un vector n-dimensional como $\sqrt{\sum_{i=1}^{n} f_i^2}$. En un espacio de Hilbert donde '*f*' es una función tenemos:

$$Longitud\, f = \sqrt{\int_a^b f^2(x)\,dx} \qquad (10)$$

En (5) se define el producto escalar de dos vectores n-dimensionales '*f*' y '*g*' como:

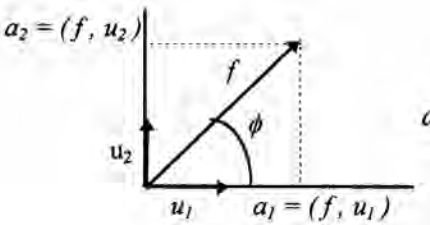
$$(f, g) = \sqrt{\sum_{i=1}^{n} f_i^2}\ \sqrt{\sum_{i=1}^{n} g_i^2}\ cos(\alpha)$$

Traspasando la ecuación anterior al espacio infinito-dimensional :

$$(f, g) = \sqrt{\int_a^b f^2(x)\,dx}\ \sqrt{\int_a^b g^2(x)\,dx}\ cos(\alpha) \quad (11)$$

Análogamente, la expresión (4) $\left((f,g) = \sum_{i=1}^{n} f_i g_i\right)$ se convierte a su equivalente en espacios de Hilbert: $(f,g) = \int_a^b f(x)g(x)dx$ *(12)*

De forma similar a lo que ocurría en los espacios n-dimensionales, en un espacio de Hilbert las funciones cuyo producto escalar es cero son ortogonales. En los espacios n-dimensionales también se cumple la siguiente propiedad: si tomamos '*n*' vectores arbitrarios 'u_i', perpendiculares entre sí y de longitud unidad, todo vector n-dimensional se puede caracterizar mediante sus componentes halladas realizando la proyección del vector sobre cada uno de los '*n*' ejes determinados por los vectores 'u_i'.



$$a_k = (f, u_k) \qquad k=1, 2, \ldots n \quad (13)$$

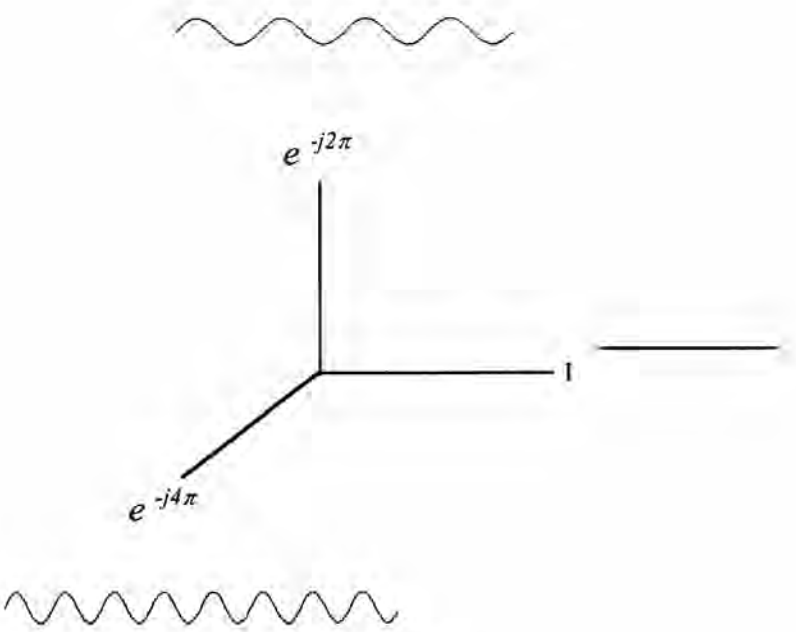
En espacios de Hilbert, un sistema de funciones $\phi_1(x)$, $\phi_2(x)$, ... $\phi_n(x)$ es ortogonal si se cumple:

$$\int_a^b \phi_j(x)\phi_k(x)dx = 0 \quad \forall\ j \neq k \quad (14).$$

La ortonormalidad del sistema se consigue cuando se cumple: $\int_a^b \phi_i^2(x)dx = 1 \quad i=1..n \quad (15)$

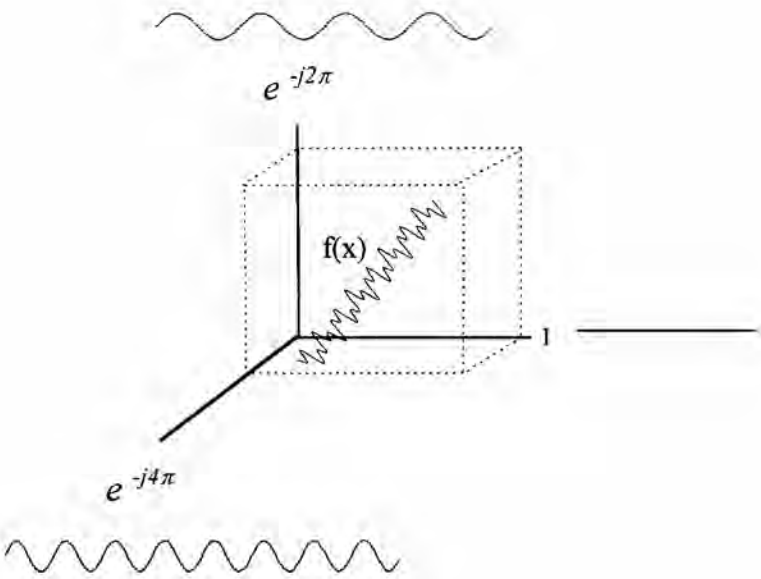
Tomando la sucesión de funciones: $1, e^{-j2\pi}, e^{-j4\pi}, e^{-j6\pi}, \ldots, e^{-j2n\pi}$

obtenemos la base $\phi_n = e^{-j2n\pi}$, $\{ e^{-j2n\pi} = cos(2n\pi) - j*sen(2n\pi) \}$ (16)



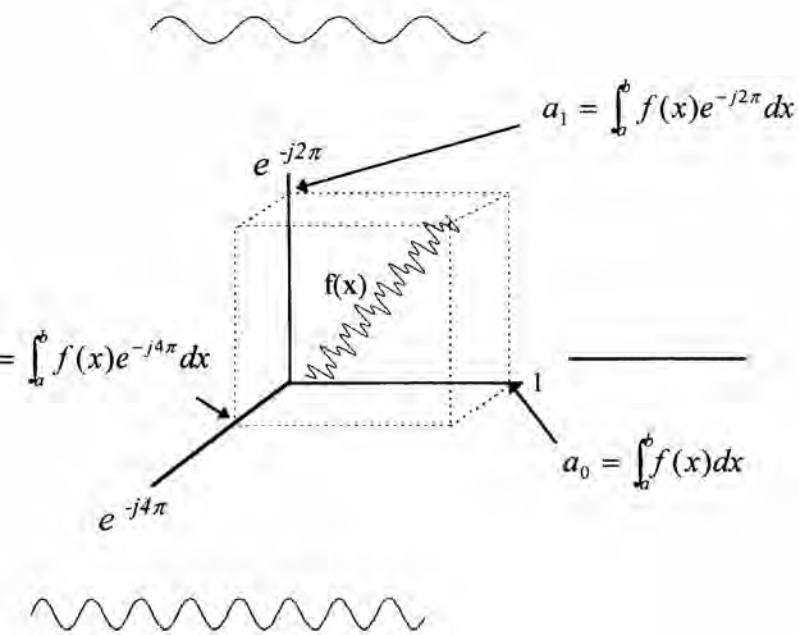
Una función f(x) se puede expresar como:

$$f(x) = a_1\phi_1(x) + a_2\phi_2(x) + \cdots + a_n\phi_n(x) \quad (17)$$



Aplicando (12) y (13) obtenemos:

$$a_n = \int_a^b f(x)\phi_n(x)dx \qquad (18)$$



El gráfico anterior ilustra la descomposición de una señal compleja empleando una base de senos y cosenos de diferentes frecuencias. El producto escalar de la función analizada por cada elemento de la base nos indica la proporción en la que cada frecuencia *(2nπ)* participa en la composición de la señal original.

Cuando existe una marcada ortogonalidad entre la señal analizada y uno de los elementos de la base *(ϕ_k)*, significa que la frecuencia determinada por *'ϕ_k'* no interviene de forma significativa en la constitución de la señal original. Analíticamente, esta característica se refleja en el término *cos(α)* que existe en la ecuación (11). Gráficamente, los términos *'a_n'* se corresponden con las proyecciones de la onda analizada sobre cada uno de los ejes determinados por los elementos de la base.

Combinando los resultados obtenidos en (16) y (18), los valores *'a_n'* se pueden expresar como:

$$a_n = \int_a^b f(x)e^{-j2n\pi}dx \qquad (19)$$

Trabajando con *'N'* señales discretas *m(kT)* cuyo recorrido se abarca desde un índice *k=0* hasta *k=N-1*, la ecuación (19) se convierte en la utilizada en el apartado anterior para ilustrar el funcionamiento de la transformada de Fourier:

$$F\left(\frac{n}{NT}\right) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT)\, e^{-j\frac{2\pi nk}{N}} \qquad (20)$$

Empleando el formalismo que nos brindan los espacios de Hilbert [ALE76], se ha llegado con sencillez a la expresión de la transformada de Fourier, cuya versión discreta nos permite realizar un análisis intuitivo de su funcionamiento y propiedades básicas.

3.2.2.3 ANÁLISIS LPC

Entre las ventajas que presenta el método de predicción lineal se encuentran [RAB93]:

- LPC proporciona un modelo adecuado de la señal de voz y sus parámetros se ajustan a las características del tracto vocal, especialmente en los sonidos sonoros del habla cuyas propiedades se aproximan más a la señal estacionaria que en los sonidos sordos [RAN95].

- Los parámetros obtenidos mediante predicción lineal muestran un espectro suavizado que proporciona la información más representativa de la voz. Esto evita perderse en los detalles ofrecidos por la transformación de Fourier.

- LPC es un método preciso, muy adecuado para computación, tanto por su sencillez como por la rapidez de ejecución que presentan algunos de los algoritmos hallados.

- Las pruebas realizadas con este método muestran muy buenos resultados en diversos campos del tratamiento automático de la voz.

DESARROLLO

El concepto básico de partida se basa en la minimización del error producido al extrapolar el valor de una muestra de voz *'x(n)'* partiendo de la información proporcionada por las *'k'* muestras anteriores *'x(n-1), x(n-2), ... x(n-k)'*, para ello utilizaremos métodos lineales:

$$\tilde{x}(n) = \sum_{i=1}^{k} a_i * x(n-i) \qquad (1)$$

Para calcular los coeficientes a_i minimizando el error, se aplican mínimos cuadrados. Primero se forma el error cuadrático medio en el intervalo de *'n'* que se desea considerar (de un máximo de *'N'* muestras):

$$L = \sum_{n} e^2(n) = x(n) - \sum_{i=1}^{k} a_i * x(n-i) \qquad 0 \le n \le N\text{-}1 \qquad (2)$$

Para obtener el valor mínimo de L, se deriva respecto a cada una de las variables $a_j \;/\; 1 \le j \le k$

$$\frac{\partial L}{\partial a_j} = 0 \qquad 1 \le j \le k \qquad (3)$$

$$\frac{\partial L}{\partial a_j} = \frac{\partial}{\partial a_j} \sum_n \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right]^2 = 0 \qquad (4)$$

$$\frac{\partial L}{\partial a_j} = \sum_n \frac{\partial}{\partial a_j} \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right]^2 = 0 \qquad (5)$$

$$\frac{\partial L}{\partial a_j} = \sum_n 2 * \frac{\partial}{\partial a_j} \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] * \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] = 0 \qquad (6)$$

$$\frac{\partial L}{\partial a_j} = -2 * \sum_n x(n-j) * \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] = -2 * \sum_n x(n-j) * e(n) = 0$$

$$\frac{\partial L}{\partial a_j} = \sum_n x(n-j) * e(n) = 0 \qquad 1 \le j \le k \qquad (8)$$

La expresión anterior se puede desarrollar como:

$$\sum_n x(n-j) * \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] = \qquad (9)$$

$$\sum_n x(n-j) * x(n) - \sum_{i=1}^{k} a_i * \sum_n x(n-j) * x(n-i) = \qquad (10)$$

$$\boxed{C_{j0} - \sum_{i=1}^{k} a_i * C_{ji}} \quad \text{donde se ha definido: } C_{ji} = \sum_n x(n-j) * x(n-i) \qquad (11)$$

MÉTODO DE AUTOCORRELACIÓN

Una manera sencilla de definir los límites de 'n' en el sumatorio, es suponer que el valor de las muestras de voz se anula fuera del intervalo $0 \leq n \leq N-1$. Lo que es equivalente a aplicar una ventana rectangular en el intervalo considerado, de esta manera:

$$\sum_n x(n-i)*x(n-j) = \sum_n x(n)*x(n+|i-j|) = r_{|i-j|} \qquad (12)$$

con lo que: $C_{ij} = C_{ji} = r_{|i-j|}$, donde los $r_{|i-j|}$ son los coeficientes de correlación de la matriz de autocorrelación. Las propiedades de esta matriz son su simetría y el hecho de que todos los elementos de la diagonal son iguales. Esto la convierte en una matriz tipo Toeplitz.

$$\sum_{i=1}^{k} r_n(|j-i|)*a_i = r_n(j), \qquad 1 \leq j \leq k \qquad (13)$$

Expresado en forma matricial:

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \cdots & r_n(k-1) \\ r_n(1) & r_n(0) & r_n(1) & \cdots & r_n(k-2) \\ r_n(2) & r_n(1) & r_n(0) & \cdots & r_n(k-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(k-1) & r_n(k-2) & r_n(k-3) & \cdots & r_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(k) \end{bmatrix} \qquad (14)$$

Esta matriz puede ser resuelta aplicando diferentes métodos, entre los más conocidos están la descomposición de Cholesky [RAB78] y el algoritmo de **Levinson-Durbin** [RAB93]:

$$L^{(0)} = r(0)$$

$$a_i^{(i)} = \frac{\left[r(i) - \sum_{j=1}^{M-1} a_j^{(i-1)} * r(|i-j|) \right]}{E^{(i-1)}}, \qquad 1 \leq i \leq k$$

$$a_j^{(i)} = a_j^{(i-1)} - a_i^{(i)} * a_{i-j}^{(i-1)}$$

$$L^{(i)} = (1 - (a_i^i)^2) * L^{(i-1)}$$

ALGORITMO DE LEVINSON-DURBIN

Estas ecuaciones se resuelven iterativamente para $i=1, 2, \ldots k$.

M indica el número de marcos escogidos en la señal analizada.

Los índices entre paréntesis indican la iteración actual o la anterior.

A los elementos de pivotaje $a_i^{(i)}$ se les denomina coeficientes PARCOR.

Este método es bastante lento, y es debido al cálculo de la matriz R_k. Al tener que calcular el coeficiente de autocorrelación $k+1$ veces se ralentiza mucho el desarrollo.

ALGORITMO DE CELOSÍA ADAPTATIVA (CAG)

Partiendo de las ecuaciones anteriores y de:

$$e(n)=x(n)-\sum_{i=1}^{k} a_i(n)*x(n-i) \qquad \text{Filtro predictor de error} \qquad (15)$$

$$f_k(n)=x(n-k)-\sum_{i=0}^{k-1} a_{k-i}(n-k)*x(n-i) \qquad \text{Filtro postdictor de error} \qquad (16)$$

obtenemos:

$$e_k(n)=e_{k-1}(n)+h_k * f_{k-1}(n-1) \qquad (17)$$

$$f_k(n)=f_{k-1}(n-1)+h_k * e_k(n) \qquad (18)$$

donde h_k es el coeficiente de reflexión o de CORrelación PARcial (PARCOR), cuyo valor es

$$h_k = - a_k^k . \qquad (19)$$

Utilizando la siguiente ecuación (usada en el método de Levinson Durbin):

$$a_i^{(i)} = \frac{\left[r(i) - \sum_{j=1}^{M-1} a_j^{(i-1)} * r(|i-j|) \right]}{E^{(i-1)}} \qquad (20)$$

se puede obtener una expresión alternativa al coeficiente PARCOR, la cual sólo hace uso de los errores directo e inverso:

$$h_{dk} = -\frac{\sum_n e_{k-1}(n) * f_{k-1}(n-1)}{\sum_n f_{k-1}^2(n-1)} \qquad (21)$$

$$h_{ik} = -\frac{\sum_n e_{k-1}(n) * f_{k-1}(n-1)}{\sum_n e_{k-1}^2(n-1)} \qquad (22)$$

Donde h_{dk} es el coeficiente directo y h_{ik} es el coeficiente inverso. Si el proceso es estacionario, ambos coeficientes tomarán el mismo valor y las normas de los dos vectores error serán las mismas. Si no es así, se puede utilizar el estimador que se obtiene de realizar la media geométrica de h_{dk} y h_{ik}, también conocido como método de Itakura, cuyas propiedades son interesantes para este caso:

$$h_{lk} = -\frac{\sum_n e_{k-1}(n) * f_{k-1}(n-1)}{\sqrt{\sum_n e_{k-1}^2(n) * \sum_n f_{k-1}^2(n-1)}} \qquad (23)$$

Con estas expresiones se pueden calcular los coeficientes de forma adaptativa, para ello se toman estimadores de los errores, los cuales se pueden actualizar en cualquier punto espacial o temporal realizando un sencillo cálculo.

El valor $\sum_n e_{k-1}(n) * f_{k-1}(n-1)$ correspondiente al numerador se estima mediante:

$$T_{k-1}(n) = \mu\ T_{k-1}(n-1) + 2\, e_{k-1}(n) * f_{k-1}(n-1) \qquad (24)$$

Los valores $\sum_n e_{k-1}^2(n)$ y $\sum_n f_{k-1}^2(n-1)$ del denominador se estiman con:

$$L_{k-1}(n) = \mu\ L_{k-1}(n-1) + e_{k-1}^2(n) + f_{k-1}^2(n-1) \qquad (25)$$

donde $0 \leq \mu \leq 1$. Ahora el coeficiente queda:

$$h_k(n+1) = -\frac{T_{k-1}(n)}{L_{k-1}(n)} \qquad 1 \leq k \leq K \qquad (26)$$

Para poder determinar de forma recursiva el valor de $h_k(n+1)$, variamos la expresión de la siguiente manera:

$$h_k(n+1)=h_k(n)+\delta_k(n) \qquad (27)$$

donde $\delta_k(n)=h_k(n+1)-h_k(n)=\frac{T_{k-1}(n)}{L_{k-1}(n)}-\frac{T_{k-1}(n-1)}{L_{k-1}(n-1)}=$ (28)

$$=-\frac{1}{L_{k-1}(n)}\left[T_{k-1}(n)-\frac{T_{k-1}(n-1)*L_{k-1}(n)}{L_{k-1}(n-1)}\right]= \qquad (29)$$

$$=-\frac{1}{L_{k-1}(n)}\left\{\left[\mu T_{k-1}(n-1)+2e_{k-1}(n)*f_{k-1}(n-1)\right]-\frac{T_{k-1}(n-1)}{L_{k-1}(n-1)}*\left[\mu L_{k-1}(n-1)+e_{k-1}^2(n)+f_{k-1}^2(n-1)\right]\right\}$$

$$=-\frac{1}{L_{k-1}(n)}\left\{2e_{k-1}(n)*f_{k-1}(n-1)+h_k(n)\left[e_{k-1}^2(n)+f_{k-1}^2(n-1)\right]\right\}= \qquad (31)$$

$$=-\frac{1}{L_{k-1}(n)}\left\{e_{k-1}(n)\left[f_{k-1}(n-1)+h_k(n)*e_{k-1}(n)\right]+f_{k-1}(n-1)\left[e_{k-1}(n)+h_k(n)*f_{k-1}(n-1)\right]\right\}=$$

$$=-\frac{1}{L_{k-1}(n)}\left[e_{k-1}(n)*f_k(n)+f_{k-1}(n-1)*e_k(n)\right] \qquad (33)$$

Con lo que la expresión recursiva buscada queda como:

$$h_k(n+1)=h_k(n)-\frac{1}{L_{k-1}(n)}\left[e_{k-1}(n)*f_k(n)+f_{k-1}(n-1)*e_k(n)\right] \qquad (34)$$

Ahora se puede escribir el algoritmo CAG de la siguiente manera:

Inicialización:

$h_i(0) = 0$; $L_{i-1}(0) = 0$; $1 \le i \le K$

Bucle principal: $0 \le n \le N-1$

$e_0(n) = f_0(n) = x(n)$

Bucle anidado: $1 \le k \le K$

$$h_k(n+1)=h_k(n)-\frac{1}{L_{k-1}(n)}\left[e_{k-1}(n)*f_k(n)+f_{k-1}(n-1)*e_k(n)\right]$$

$$e_k(n+1) = e_{k-1}(n+1) + h_k(n+1)*f_{k-1}(n)$$

$$f_k(n+1) = f_{k-1}(n) + h_k(n+1)*e_k(n+1)$$

$$L_{k-1}(n+1) = \mu\, L_{k-1}(n) + e_{k-1}^2(n+1) + f_{k-1}^2(n)$$

Como se puede ver, la idea es ir actualizando todos los coeficientes en cada muestra, de forma que, tras 'N' muestras, se tengan unos coeficientes que representen al segmento que se está estudiando.

El número de coeficientes, 'K', que se emplea ha de ser pequeño, para tardar poco en los cálculos, pero lo suficientemente grande como para que la función de transferencia devuelva buenas aproximaciones. Generalmente 'K' suele oscilar entre 10 y 20.

Los valores de 'μ' se toman habitualmente cercanos a 1. Su función es la de estabilizar el proceso.

A diferencia del algoritmo de Levinson-Durbin cuyos coeficientes se pueden pasar directamente a la función de transferencia, los valores obtenidos tras el procesamiento de la celosía de gradiente adaptativo requieren de una fase de adaptación.

Los valores que tenemos son los coeficientes de pivotaje $a_i^{(i)}$, para obtener los $a_j^{(i)}$ buscados, aplicamos la conversión descrita en el algoritmo de Levinson-Durbin.

Algoritmicamente, este paso se puede realizar de la siguiente manera:

Inicialización:	$1 \le j \le K$	
$a_i(j) = 0$		
Bucle principal:	$1 \le k \le K$	
$a_i(0) = -1$		
Bucle interno:	$c(i) = a_i(i) + a(k)*a_i(k-i)$	$0 \le i \le K+1$
Bucle interno:	$a_i(i) = c(i)$	$0 \le i \le K+1$

ALGORITMO DE ADAPTACIÓN CAG CON LEVINSON-DURBIN

FUNCIÓN DE TRANSFERENCIA

Una vez que se han obtenido los coeficientes, el último paso es el de obtener la característica espectral que se desea usando estos valores. La función de transferencia que permite calcular el espectrograma de la voz es:

$$H(f) = \frac{G}{1 - \sum_{k=1}^{K} a_k * e^{\frac{-2\Pi ikf}{2M}}} \qquad (35)$$

donde *'G'* es la ganancia (cuyo valor se suele fijar en la unidad) y *'M'* la cantidad de resultados (definición frecuencial) que se quieren obtener, siendo *'f'* el dato que se está calculando.

Una posible implementación para obtener las frecuencias correspondientes a un conjunto de valores a_k es:

Bucle principal: $0 \leq m < M$

Inicialización:

Parte real $\Leftarrow 1$

Parte imaginaria $\Leftarrow 0$

Bucle interno: $1 \leq k < K$

Parte real $\Leftarrow$ Parte real - a_k * Coseno $\left(\frac{mk\Pi}{M}\right)$ * Radio^k

Parte imaginaria $\Leftarrow$ Parte imaginaria + a_k * Seno $\left(\frac{mk\Pi}{M}\right)$ * Radio^k

Resultado:

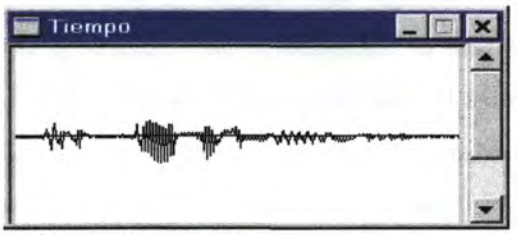
Valor frecuencial $\Leftarrow L_{10}\left(\frac{1}{\sqrt{\left[(Parte\,real)^2 + (Parte\,imag)^2\right]}}\right)$

ALGORITMO DE FUNCIÓN DE TRANSFERENCIA

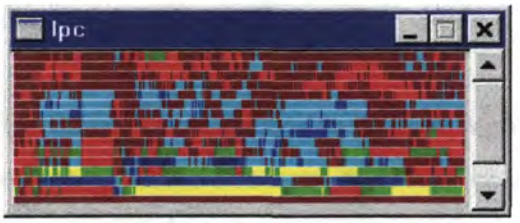
Radio representa el círculo de búsqueda en el plano complejo, se suele utilizar el círculo unidad (Radio=1), aunque existen algoritmos como el 'chirp z-Transform' [RAB70], que realizan búsquedas dentro del plano complejo para hallar polos muy cercanos que se confunden entre sí.

En las siguientes figuras se representan visualmente los coeficientes LPC y el espectro obtenido tras aplicar la función de transferencia sobre los mismos. Las muestras pertenecen a la grabación de la palabra 'economía'. Los parámetros empleados son:

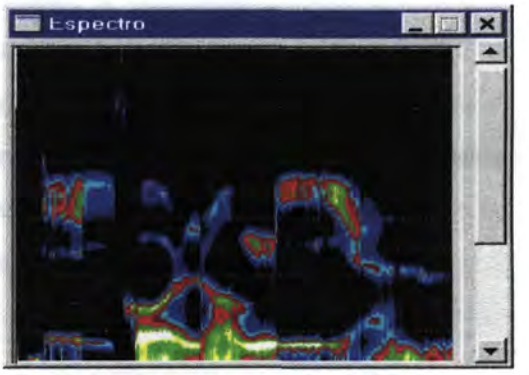
K = 16, M = 256, Radio = 1, μ = 0.98, Frecuencia de muestreo = 11025 muestras/sg., Se computa una muestra de cada 50.



SEÑAL EN EL TIEMPO



PARÁMETROS LPC



ESPECTRO DE VOZ OBTENIDO TRAS APLICAR LA FUNCIÓN DE TRANSFERENCIA

3.3. FONÉTICA ACÚSTICA ESPAÑOLA

En este apartado se pretende ofrecer un repaso de las características acústicas más relevantes de los principales sonidos del español, para ello se tomará como referencia la clasificación y estudios aportados por Antonio Quilis en [QUI93].

Para ilustrar los aspectos más significativos de cada sonido se emplearán diversos espectros de voz producidos informáticamente, el método utilizado para producir estos espectros será el de predicción lineal implementado con el filtro de celosía adaptativa. Aunque la transformada de Fourier sería más apropiada para observar fenómenos puntuales como barras de oclusión, vibraciones, etc. se ha preferido usar LPC para poder así contrastar más adecuadamente las mejoras que se consigan a lo largo de la tesis en la representación espectral basada en este tipo de algoritmo.

La descripción de sonidos que aquí se presenta no tiene una relación directa con los desarrollos que se realizarán a lo largo de la tesis en el campo de la fonética acústica, sin embargo, se ha considerado adecuado aportar una base mínima de conocimientos en la materia, que ayuden a la comprensión del resto de las investigaciones.

La clasificación de sonidos que se repasará figura en la tabla siguiente:

CONSONANTES

	Bilabial		Labiodental		Dental		Interdental		Alveolar		Palatal		Velar	
	sordas	sonor.	sordas	sonor.	sordas	sonor	sordas	sonor	sordas	sonor	sord.	sonor	sord	sonor
Oclusivas	[p]	[b]			[t]	[d]							[k]	[g]
Fricativas		[β]	[f]				[θ]	[d.]	[s]			[j]	[x]	[γ]
Africadas											[t∫]	[dξ]		
Nasales		[m]		[μ]		[n,]		[n.]		[n]		[n_{μ}][n_{δ}]		[ɳ]
Laterales						[l,]		[l.]		[l]		[λ][l_{δ}]		
Vibr.simple										[r]				
Vibr. doble										[rr]				

VOCALES

	i	e	a	o	u
Orales	[i]	[e]	[a]	[o]	[u]
Nasales	[i_{φ}]	[e_{φ}]	[a_{φ}]	[o_{φ}]	[u_{φ}]

Los primeros sonidos que se analizarán son los vocálicos, cuya característica principal se basa en la existencia de estructuras de formantes claras, fruto de la emisión del flujo de aire por el conducto bucal sin apenas resistencia, y con las cavidades resonadoras potenciando los armónicos distintivos de cada vocal.

Los sonidos vocálicos que analizaremos serán los correspondientes a los fonemas /i/, /e/, /a/, /o/, /u/ desdoblados en alófonos orales y nasales. Un alófono nasal se produce cuando su correspondiente fonema vocálico se encuentra entre una pausa y una consonante nasal o entre dos consonantes nasales.

En el gráfico superior de la figura 3.10 se presentan los cinco sonidos vocálicos orales, incluidos en las palabras:

[i] pipa [e] pepa [a] papa [o] popa [u] pupa

y los cinco sonidos vocálicos nasales, incluidos en las palabras:

[i_{φ}] mimo [e_{φ}] memo [a_{φ}] mama [o_{φ}] mono [u_{φ}] mundo

Los recuadros indican las zonas de los espectros donde se encuentran los sonidos estudiados.

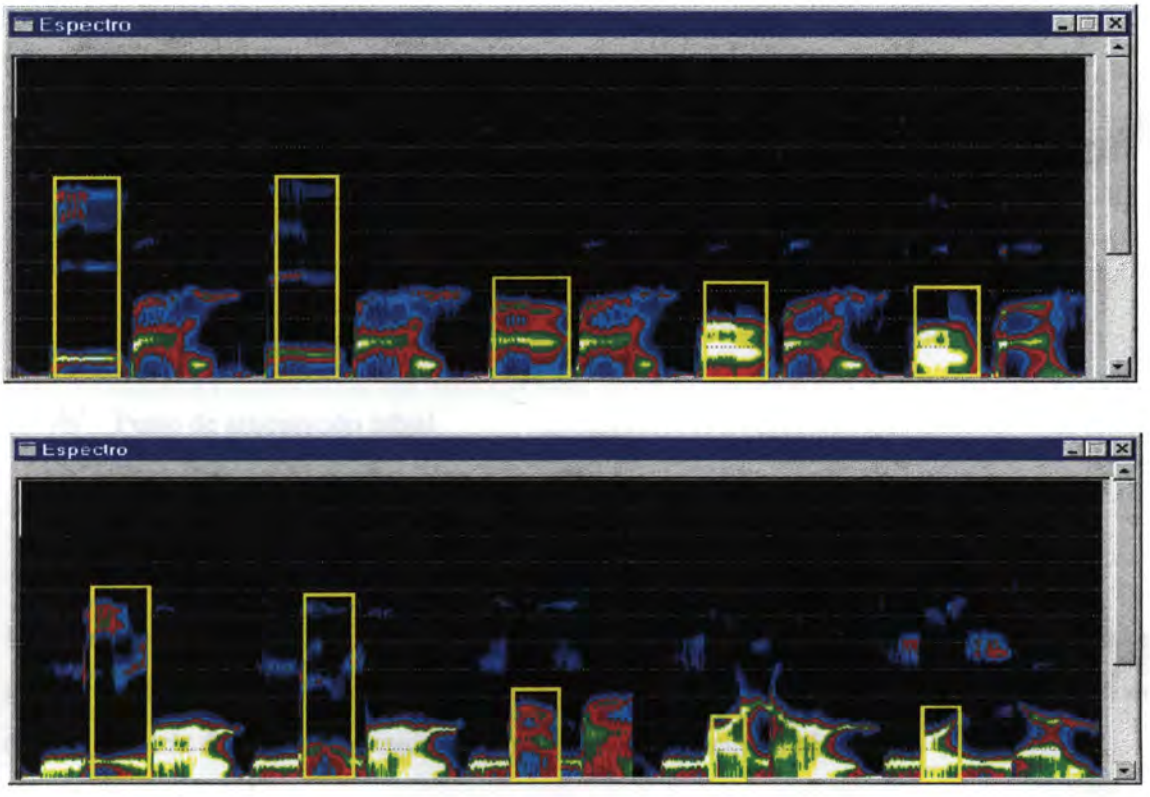
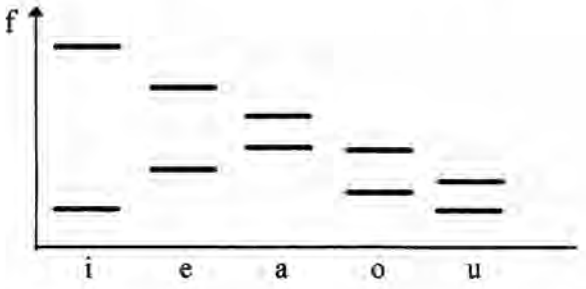


Figura 3.10
Ejemplos de los sonidos vocálicos orales y nasales.
Espectro superior: pipa, pepa, papa, popa, pupa
Espectro inferior: mimo, memo, mamo, mono, mundo

Las vocales presentan estructuras de formantes bien definidas. Su situación exacta varía según el hablante y la realización del habla. Las posiciones relativas de los dos primeros formantes son:



En el caso de las vocales nasales, la intensidad del primer formante se reduce.

El siguiente grupo que se analizará será el de las consonantes oclusivas (explosivas). Este tipo de sonidos surge del cierre u oclusión de los órganos fonadores durante un intervalo de tiempo, seguido de su abertura con la consiguiente salida brusca de aire (explosión).

Los fonemas básicos son:

- oclusivas sordas:
  - /p/ Punto de articulación labial
  - /t/ Punto de articulación dental
  - /k/ Punto de articulación velar

- oclusivas sonoras:
  - /b/ Punto de articulación labial
  - /d/ Punto de articulación dental
  - /g/ Punto de articulación velar

Los fonemas /b/ y /g/ poseen sendos alófonos fricativos [β], [γ] que se producen cuando no se encuentran detrás de pausa o de consonante nasal. /d/ posee el alófono fricativo [d.] cuando no se encuentra detrás de pausa, de consonante nasal o de consonante labial.

La figura 3.11 contiene los sonidos oclusivos y fricativos obtenidos en las siguientes palabras:

[p] **p**asa	[t] **t**asa	[k] **c**asa
[b] **b**ado	[d] **d**ato	[g] **g**ato
[β] a**v**aro	[d.] a**d**oro	[γ] a**g**arro

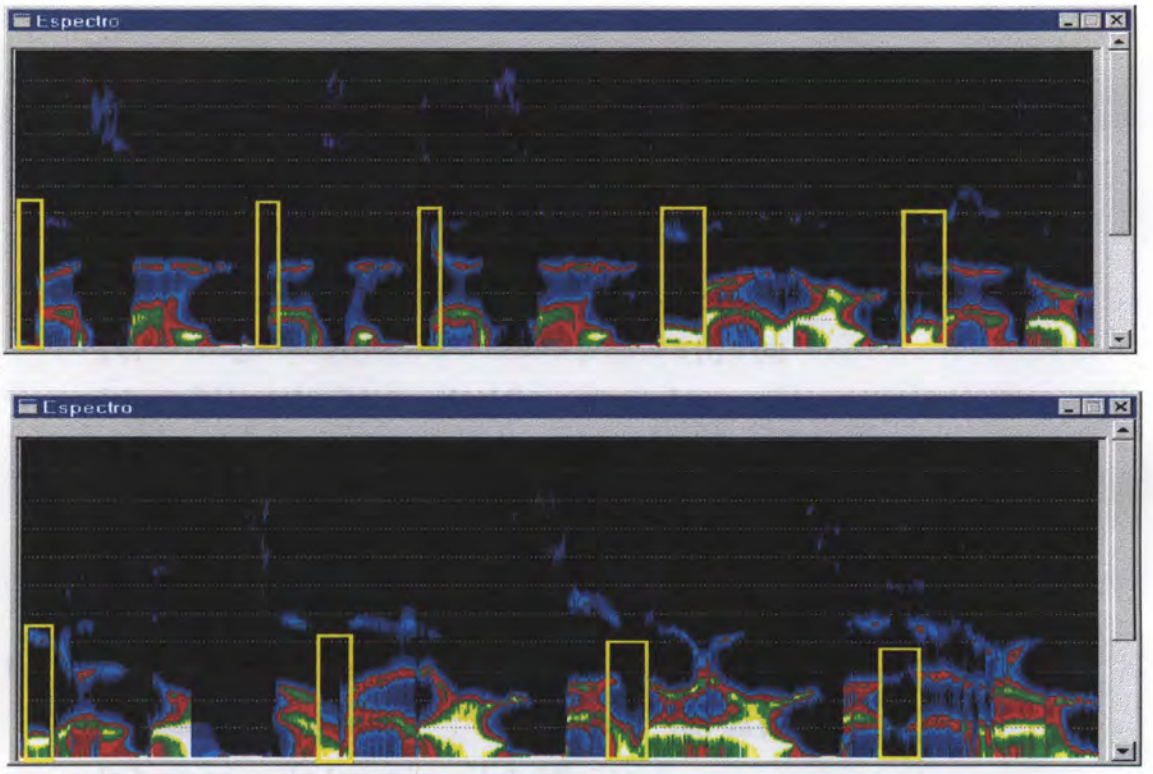


Figura 3.11

Ejemplos de sonidos oclusivos y de fricativos sonoros.

Espectro superior: pasa, tasa, casa, bado, dato

Espectro inferior: gato, avaro, adoro, agarro

Los sonidos oclusivos sordos presentan una zona de silencio seguida por una breve barra de explosión vertical, que tiene mayor duración temporal en el sonido [k]. La barra de explosión contiene más energía en la zona baja del espectro en el caso bilabial, en la zona media cuando se trata del sonido [t] y en la parte alta para [k].

Los sonidos fricativos sonoros [β], [d.], [γ] poseen la característica de sonoridad representada por una frecuencia muy baja (barra de sonoridad) producida por la vibración básica de las cuerdas vocales (en los espectros que aquí se muestran la barra de sonoridad ha sido filtrada).

A continuación analizaremos las consonantes nasales, en ellas se produce un cierre de los órganos articulatorios bucales, con la consiguiente expulsión del aire a través de los conductos nasales.

Los fonemas nasales son:

/m/ Punto de articulación labial

/n/ Punto de articulación alveolar

/ñ/ Punto de articulación palatal

Los alófonos más comunes son:

[m] Bilabial, sonido del fonema /m/. También cuando el fonema /n/ precede a una consonante labial [p], [b] o [m].

[μ] Labiodental, cuando el fonema precede a una [f]

[n.] Linguointerdental, cuando el fonema precede a [θ]

[n,] Linguodental, cuando el fonema precede a [t] o [d]

[n] Linguoalveolar, cuando el fonema precede a vocal, consonante alveolar o pausa

[n_{μ}] Linguopalatalizada, cuando el fonema precede a una consonante palatal

[η] Linguovelar, sonido del fonema /ñ/. También cuando el fonema precede a una consonante velar, [k], [g] o [x]

La figura 3.12 muestra espectros de estos sonidos, contenidos en las palabras:

[m] **mamá** [n] **un** loro [μ] **un** farol [n,] **un** tomo

[n.] **un** cero [n_{μ}] **un** chico [η] **caña**

El primer formante nasal aparece mucho más alto que la barra de sonoridad de otras consonantes, y esta es una buena indicación de la nasalidad.

El siguiente grupo de consonantes que se presenta es el de las fricativas. Estos sonidos se producen cuando se realiza un estrechamiento entre dos órganos articulatorios produciéndose la fricación.

Existen cinco fonemas fricativos: /f/, /θ/, /s/, /j/ y /x/. A cada fonema le corresponde un sonido, salvo /j/ que presenta dos alófonos:

[dξ] Africado, cuando se encuentra después de pausa, de consonante nasal, o de [l]

[j] Aparece en el resto de los casos

Las realizaciones fricativas de los fonemas oclusivos pueden clasificarse también en este grupo de sonidos.

En castellano existe un fonema africado, que acústicamente se compone de la secuencia oclusivo+fricativo. Los alófonos existentes son: [dξ]y [j] que ya han sido estudiados, y [tʃ].

Un ejemplo en el que aparece el sonido [tʃ] es la palabra '**ch**ico'.

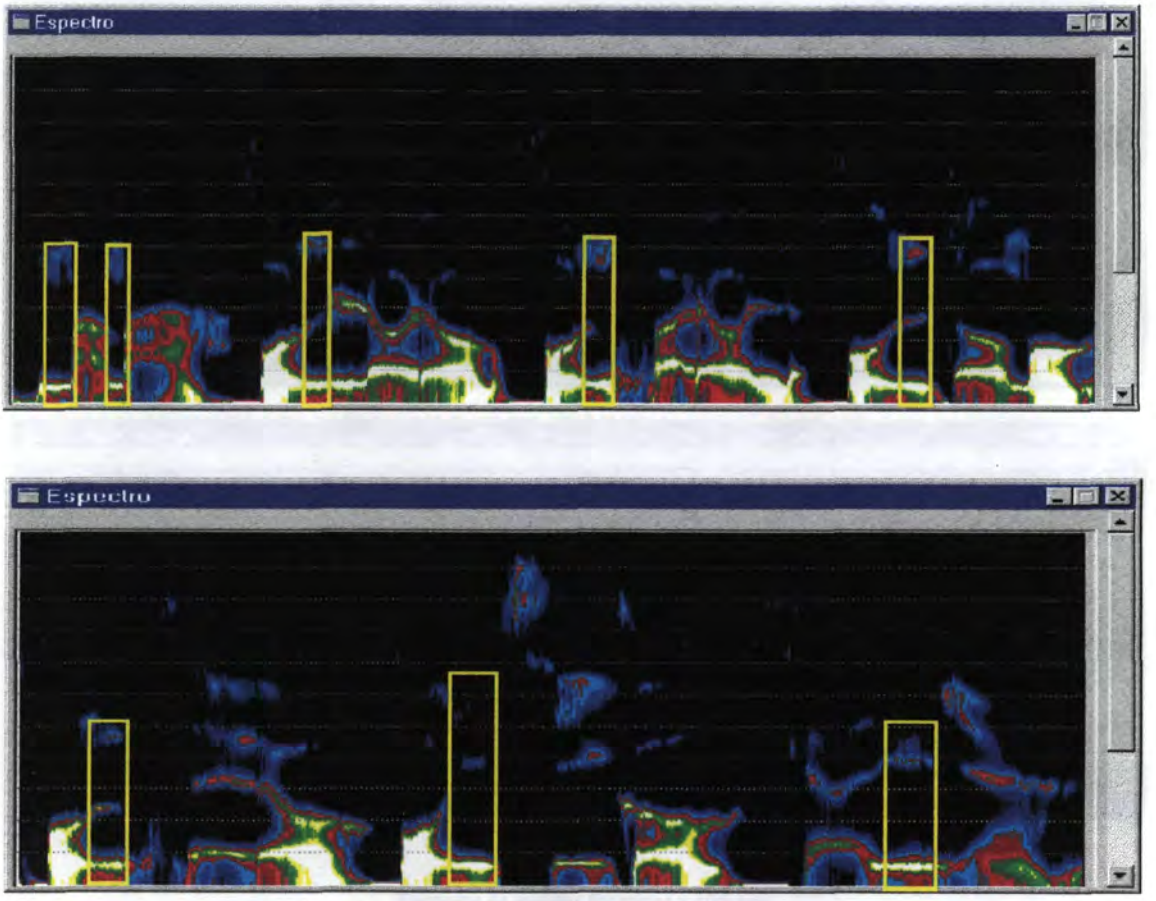


Figura 3.12
Ejemplos de los sonidos nasales.
Espectro superior: mamá, un loro, un farol, un tomo
Espectro inferior: un cero, un chico, caña

La figura 3.13 muestra la apariencia de los sonidos fricativos y africado incluidos en las palabras:

[f] ca**f**é	[θ] **z**ona	[s] ca**s**a
[x] pa**j**e	[dξ] **y**o	[j] ma**y**o

Las fricativas se diferencian de las demás consonantes por el ruido que presentan. Para distinguir las fricativas españolas entre sí, se recurre a determinar la altura frecuencial a la que se presenta su mayor energía. La mayor parte de estos sonidos poseen resonancias altas. En cuanto a intensidades, los sonidos más fuerte son los correspondientes al fonema /s/. /f/ y /θ/ presentan intensidades muy débiles.

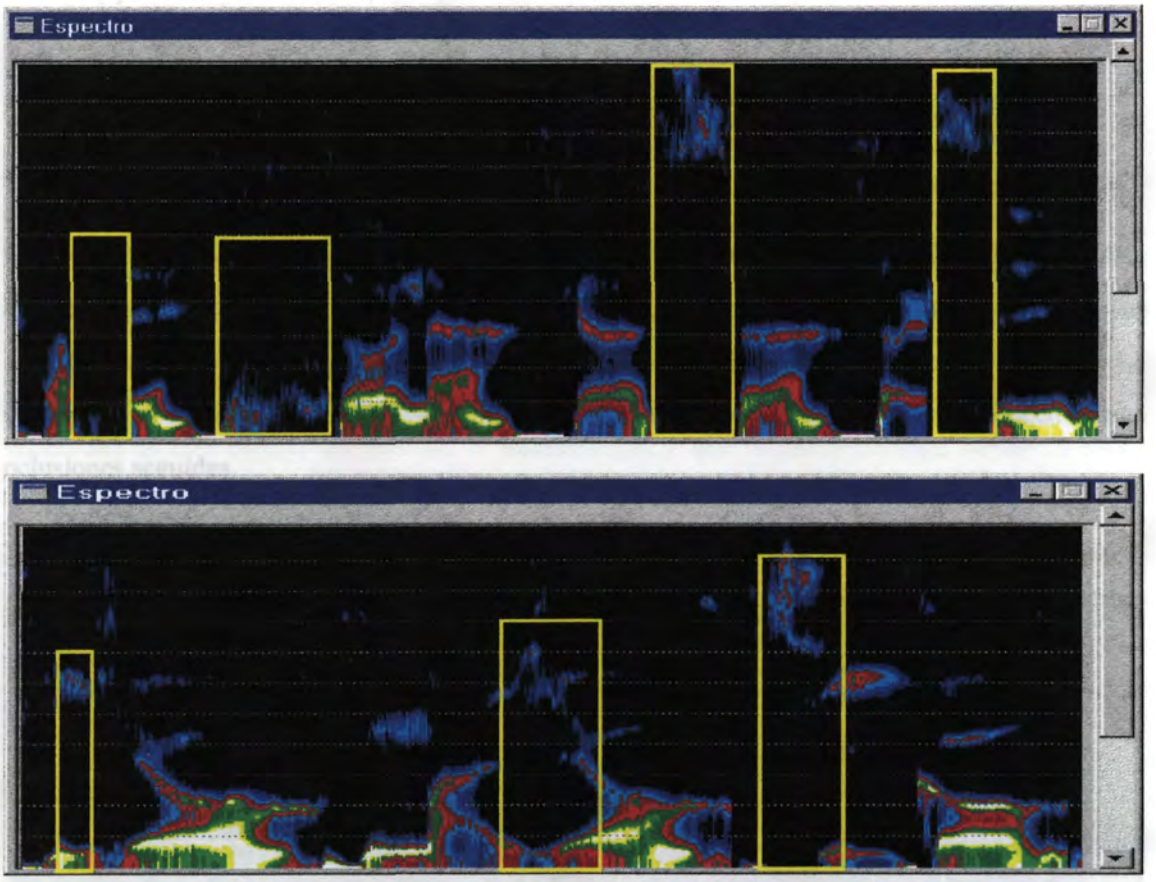


Figura 3.13
Ejemplos de sonidos fricativos y africado.
Espectro superior: café, zona, casa, paje
Espectro inferior: yo, mayo, chico

El grupo que completa el alfabeto español es el de las consonantes líquidas, que se producen al pasar el aire por la cavidad bucal con una oclusión central o lateral, de manera que estas consonantes se encuentran acústicamente entre las vocales y las demás consonantes.

Los fonemas líquidos son:

- laterales: /l/ y /λ/
- vibrantes /r/ y /rr/

alófonos de /l/:

[l.] Interdental, como en 'dulce'

[l,] Dental, como en 'toldo'

[l_{δ}] Palatal, como en 'el chico'

[l] El resto de los casos, ejemplo '**a**l**a**'

sonidos vibrantes:

[r] ca**r**o [rr] ca**rr**o

La figura 3.14 muestra los sonidos líquidos especificados. Debido a la poca resistencia a la salida del aire que existe en las consonantes laterales, acústicamente existen formantes similares a los sonidos vocálicos. Las vibrantes se producen por medio de interrupciones a la salida del aire. La vibrante simple presenta una breve oclusión, mientras que en la múltiple se producen varias oclusiones seguidas.

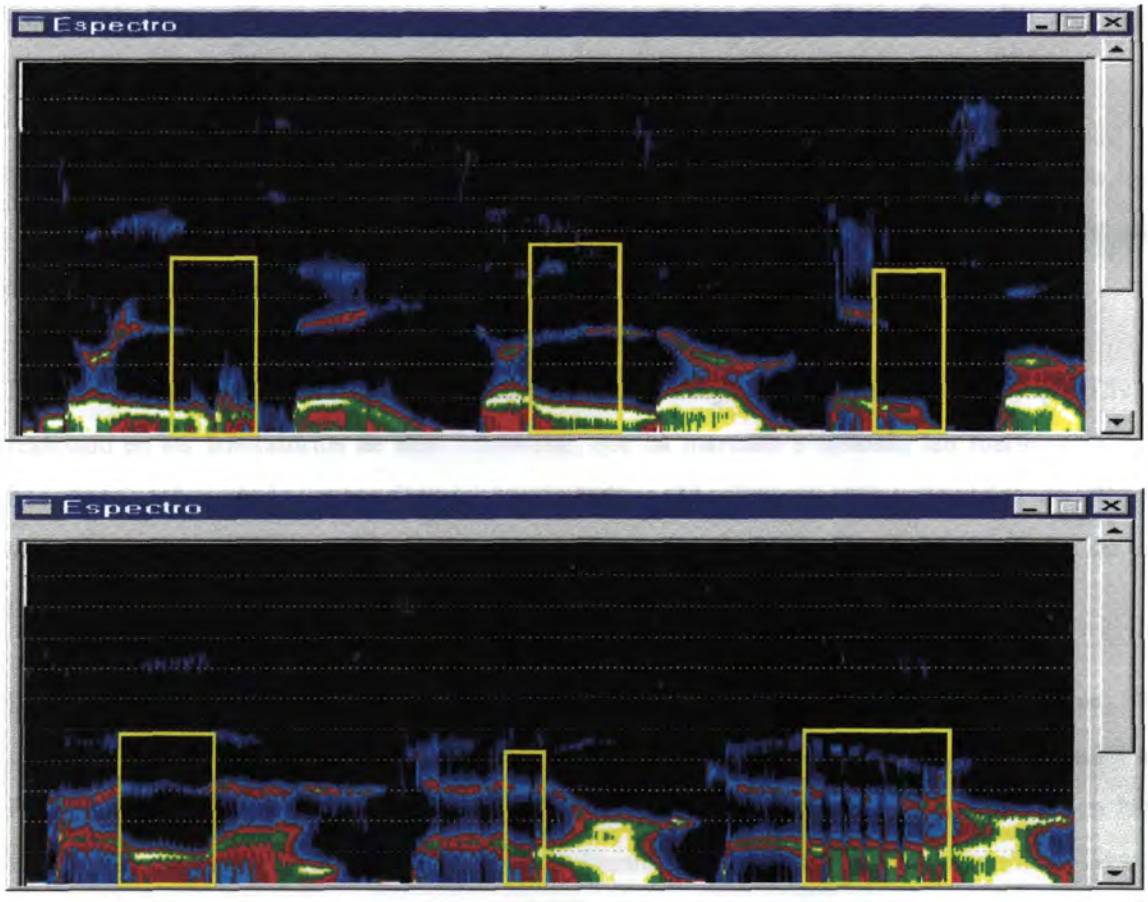


Figura 3.14
Ejemplos de los sonidos líquidos.
Espectro superior: dulce, toldo, el chico
Espectro inferior: ala, caro, carro

3.4 FONÉTICA ACÚSTICA. LÍNEAS DE INVESTIGACIÓN

En este apartado se pretende dar un breve repaso a las diferentes investigaciones desarrolladas en el campo de la fonética acústica, centrándose en aquellas publicaciones que inciden en el estudio de las características acústicas del habla como base para la realización de aplicaciones en el área del tratamiento automático de la voz.

Es importante destacar que a lo largo de los últimos 45 años, el número de publicaciones realizadas en este campo tan extenso es innumerable, y que la selección aquí mostrada incide únicamente en aquellos aspectos que se han considerado de interés para orientar la realización de la tesis hacia la consecución de objetivos de índole práctica. Para una mayor profundización en el campo, en [TOH92] se recopilan por temas aportaciones en la misma línea que las aquí presentadas.

La selección realizada pretende así mismo asegurar que la tesis tome como punto de partida los últimos avances logrados en el área, motivando nuevos enfoques y evitando la repetición de investigaciones puntuales.

El estudio con el que comenzaremos será el clásico trabajo de Peterson & Barney [PET52], realizado en los laboratorios de Bell Telephone, que ha marcado e influido tan fuertemente las investigaciones posteriores en el campo que incluso 38 años después se pueden encontrar artículos [WAT90] que no sólo lo referencian, sino que hacen de este estudio el núcleo de la publicación.

En este trabajo se grabaron 10 vocales en un contexto /hVd/ pronunciado por 33 hombres, 28 mujeres y 15 niños. Una vez procesados los datos, se obtuvieron medidas de los formantes F1 a F3 y del tono fundamental (F0). Se descubrió que existe una considerable variabilidad en las frecuencias de los formantes de los distintos hablantes y que se producen solapamientos entre ocurrencias de vocales adyacentes, a pesar de ello las vocales presentaron un alto porcentaje de aciertos en su identificación.

En la producción del habla se producen muchas variaciones debidas a la complejidad del proceso, la intervención de diferentes individuos, los cambios físicos y de estado de ánimo de un mismo hablante a lo largo del tiempo, influencias dialectales, etc.

En [PET52] se diseña un experimento que permite cuantificar y relacionar parámetros espectrales de las vocales inglesas en contextos /hVd/; así mismo, se comprueba cuáles son las vocales mejor reconocidas en pruebas de audición atendiendo a las posiciones de sus formantes.

El resultado más importante obtenido es un mapa de vocales inglesas en el plano Formante 1 / Formante 2; este tipo de mapas ha sido utilizado y mejorado repetidamente hasta nuestros días.

Este estudio tuvo una gran importancia en su época, puesto que sentó las bases de utilización de técnicas de calibración y medida en espectrógrafos, además de mostrar un método que incluye obtención de datos, aleatorización de procedimientos de medida y audición, uso de técnicas estadísticas, etc.

La base de datos utilizada en [PET52] tiene bastantes limitaciones de diseño, en primer lugar, las medidas fueron tomadas en instantes puntuales de tiempo, por lo cual no existe información de la evolución de los parámetros espectrales. Esta laguna es especialmente importante debido al papel fundamental que ejercen en el reconocimiento de voz las propiedades dinámicas del habla tales como la evolución de los formantes y la duración espectral de cada sonido.

Otras limitaciones son [HIL95]:

1.- No se tiene información de los dialectos de los hablantes que realizaron las grabaciones.
2.- Los resultados no se agruparon separadamente para hombres, mujeres y niños.
3.- No existe información de las edades de los niños empleados en el estudio y además el número de los mismos no era suficientemente grande como para realizar un adecuado estudio estadístico.
4.- No se puede determinar el origen (por sexo, edades, etc.) de las muestras empleadas.

En [HIL95] se intenta minimizar estas limitaciones, se grabaron secuencias /hVd/ pronunciadas por un numeroso grupo de hombres, mujeres y niños; se tomaron medidas de la duración de las vocales, frecuencias fundamentales y primeros formantes. Al igual que en [PET52] las señales se presentaron a un conjunto de oyentes para su identificación.

Tras realizar diversos análisis discriminantes se clasificaron las señales usando varias combinaciones de las medidas acústicas. Un resultado fundamental del estudio es la necesidad de

utilizar algo más que un simple instante de tiempo por muestra para conseguir una correcta clasificación vocálica.

Los trabajos sobre identificación de formantes forman un grueso bloque de investigación en un área que se considera fundamental para el avance en casi todos los campos del tratamiento automático de la voz, incluidos el reconocimiento y la síntesis del habla. La mayor parte del contenido de este apartado se centra en la recopilación de estudios basados en este concepto

Si bien es conocido que la altura de los formantes y de la frecuencia fundamental varía con la edad y el sexo, (la altura de los formantes producidos por los niños decrece con la edad y en las niñas es mayor que en los niños, aproximadamente un 10% superior), en [BUS95] se realiza un experimento que analiza dichas variaciones. En este estudio se pretende determinar si existen diferencias claras relacionadas con el sexo y la edad en los valores frecuenciales de los formantes producidos por niños y niñas de 5,7,9 y 11 años de edad.

Como resultados principales tenemos que los valores de F0 varían de forma inversamente proporcional al incremento de edad, sin embargo, no se aprecian diferencias significativas de este parámetro atendiendo al sexo. En el caso de la altura de F1, F2 y F3, los valores se hacen menores a medida que aumenta la edad, y las posiciones frecuenciales de las niñas son superiores a las presentadas por los niños.

En [KAT95] se obtuvo F0 a partir de las grabaciones de vocales en contextos /hVd/ producidos por 10 hombres, 10 mujeres y 30 niños con edades entre los 3 y los 7 años. Después se sintetizaron estas mismas vocales con F0 constante (sin variaciones a lo largo del tiempo) y con F0 a una altura frecuencial igual a la media obtenida en los hablantes. Tras un experimento de identificación vocálica, se obtuvo como conclusión que las variaciones de F0 no representan un factor importante en la identificación de vocales sobre sílabas aisladas.

El experimento que se realizó en [ASS95] complementa y amplía al diseñado en [KAT95]; para ello se siguen los mismos pasos, pero trabajando con los 4 primeros formantes (F1-F4). En este caso, la identificación vocálica empeora significativamente cuando se utilizan formantes sintetizados planos (sin variaciones temporales) y promediados. Estos resultados enfatizan la importancia que en el reconocimiento de la voz tiene la información que proporciona la situación y evolución de los formantes.

Con el fin de situar las vocales castellanas en el plano F1-F2, se realiza en [FER93] un estudio de dispersión de las 5 vocales creadas mediante síntesis de voz, para ello se realiza un experimento de audición, en el que en primer lugar se generan sintéticamente diversas vocales y después se valida su percepción mediante un conjunto de oyentes.

Los resultados del experimento se centran en los valores frecuenciales de las vocales halladas (superiores a los expuestos en diversas publicaciones [MAR94], [ROM88]). Como conclusión se aprecia que el campo de dispersión de las vocales es mayor desde el punto de vista perceptivo que desde el punto de vista de producción.

En [MAR90], Martínez Celdrán presenta como artículo una herramienta (programa informático) de representación visual de vectores bidimensionales, lo que viene a mostrar la falta de software de propósito general circulando entre los investigadores.

El estudio de las características espectrales de la voz se complica mucho debido a la variabilidad que se produce en el habla cuando se pronuncian varios sonidos seguidos de forma que existan interacciones entre ellos, lo cual, lejos de resultar extraño, es la situación normal del habla continua.

Una fuente importante de variación en la realización espectral de las vocales está dada por el contexto de consonantes que puede existir. A la influencia de este fenómeno habitualmente se le denomina coarticulación. Se ha desarrollado una gran cantidad de sistemas que modelizan la coarticulación, pero ninguno de ellos parece poder explicar todas las observaciones encontradas [TOK93]. Cada modelo tiene que resolver el problema de abarcar la variación de las realizaciones de las vocales aisladas respecto a los espectros producidos en segmentos de voz coarticulados.

Gran parte de la dificultad que se presenta viene dada por la variabilidad que introduce la utilización de diversos hablantes, habiéndose observado una marcada dependencia de la persona que pronuncia en los fenómenos de coarticulación, hasta el punto de que éste podría ser considerado como un parámetro de identificación del hablante [SU74]. En este estudio se descubre que la coarticulación en las nasales (especialmente en la 'm') varía significativamente con las distintas personas, y como consecuencia, este factor puede ser tomado en cuenta en el campo del reconocimiento del hablante. Por otra parte, la variación existente entre diversos

individuos en el fenómeno de coarticulación, nos hace descartar la hipótesis de universalidad en las características de articulación fonética.

Un reciente estudio en el campo se documenta en [HEU96], donde se realiza un experimento para determinar la variabilidad en el hablante que se produce en la coarticulación de las vocales holandesas 'a','i','u'. Primero se extrajeron los formantes F1 a F3 de muestras de voz coarticuladas con diferentes combinaciones de consonantes, después se efectuaron diversas operaciones estadísticas basadas en análisis lineales discriminantes, y por fin, analizando los resultados, se obtuvo la conclusión de que si bien el efecto de coarticulación ofrece una medida razonable para la correcta identificación de hablantes, no es por si solo lo suficientemente seguro como para poder emplearlo aisladamente en este fin.

En [MEM78] se realiza uno de los estudios primarios destinados a identificar las posiciones de los formantes en vocales con contextos consonánticos. [KEW95] amplía éste y otros trabajos con el objetivo de determinar los efectos de los contextos consonánticos en la discriminación de la frecuencia de los formantes.

Debido a la gran variedad de combinaciones existentes entre vocales y consonantes, la investigación se centró únicamente sobre la /i/ sintetizada aisladamente y en contextos /CVC/ con las consonantes /b,d,g,z,m,l/. La elección de la vocal /i/ no es casual, sino que se trata de uno de los dos casos más complejos (/i/, /ae/) analizados en [MEM78].

Examinando los parámetros de estudio empleados, se destacan dos factores discriminantes: la longitud de la porción estable de la vocal y la separación de F1 y F2 en las transiciones de los formantes. Ambos factores se realzan en vocales con contextos consonánticos. Otra conclusión significativa se centra en la mayor influencia que ejercen las consonantes sobre la evolución del formante F2 respecto a F1.

En [PIC95] se realiza una completísima labor de recopilación de los estudios principales realizados hasta la fecha en el campo de la caracterización fonética de las consonantes intervocálicas. La importancia del tema viene dada no sólo por la frecuencia con la que se presenta este tipo de estructuras dentro de las sílabas, sino porque también es un factor clave su influencia en la percepción del habla cuando estas consonantes se encuentran entre sílabas. Resulta importante obtener un mayor conocimiento en el campo para poder clarificar las diferentes teorías existentes en la percepción de los sonidos.

En el habla continua se producen posiciones intervocálicas de las consonantes con mayor frecuencia que en comienzos o finales de frase. La mayor parte de los estudios de fonética acústica en el área se basan en unidades de tipo sílaba como patrones de partida, esto tiene la ventaja de reducir a un tamaño razonable el número de combinaciones posibles a tratar; por otra parte, el uso de unidades silábicas o similares presenta el problema de que los principios y conclusiones hallados en su estudio no siempre son ciertos y aplicables en el habla continua.

En la percepción de las consonantes intervocálicas intervienen diferentes factores acústicos como los siguientes: variaciones en las frecuencias y anchos de banda de los formantes, segmentos de sonoridad, barras de oclusión, fricación, aspiraciones, silencios, ceros, polos, etc. En algunos casos se requiere la simultaneidad de varios de ellos para la identificación fonética, mientras que otras veces es suficiente que se presente alguno de forma aislada, por ejemplo silencio+barra de oclusión, variación de las frecuencias a lo largo del tiempo, etc.

Diferentes estudios [MAS75], [REP78] muestran la mayor importancia del segmento consonante-vocal (CV) en las ocurrencias VCV analizadas, esto se documenta claramente en [REP78] en donde se aprecia una mayor dificultad para la discriminación VC que en los segmentos CV. Sólo en los casos en los que las pistas proporcionadas en CV no son suficientes, se resuelve la ambigüedad mediante la porción VC.

Las transiciones vocálicas ofrecen una información vital en la localización del punto de articulación consonántico, en [MAR94] se expone que las consonantes se perfilan por las transiciones de las vocales adyacentes y en el caso de las oclusivas, por la altura de la mayor intensidad en la barra de explosión.

Buena parte de los estudios realizados sobre la evolución de los formantes, determinan la posición aproximada de los diferentes locus, sin precisar las diferencias existentes entre las vocales posteriores y anteriores a las consonantes estudiadas. El objetivo de [MOR90] es el de precisar el papel que cumplen las transiciones de las vocales anterior y posterior a una consonante en la identificación del punto de articulación de dicha consonante, para ello, y mediante el uso de un espectrógrafo se generaron palabras de estudio conteniendo consonantes oclusivas sordas y se realizó un experimento auditivo destinado a determinar cuál de las transiciones vocálicas tenía más importancia para predecir el punto de articulación de la oclusiva intervocálica.

Los resultados de [MOR90] muestran con mucha claridad el papel primordial de las transiciones de los formantes de la vocal posterior a la consonante oclusiva estudiada, mientras que las transiciones de los formantes de la vocal anterior no aparecen como representativos para el fin perseguido: la identificación del punto de articulación de las consonantes oclusivas sordas.

En la identificación de consonantes oclusivas, existen varios factores que proporcionan pistas para hallar el lugar de articulación. Las primeras investigaciones se centraron en características individuales de la señal de voz, tales como las barras de oclusión y la transición de formantes [BLU79], [BLU80].

Investigaciones más recientes mantienen la importancia del análisis espectral enfocado en la región de cambio entre el fin de la barra de oclusión y el comienzo de los formantes de la siguiente vocal. Las propiedades espectrales de las barras de oclusión y la evolución completa de los formantes forman pistas secundarias, su función es incrementar la efectividad del reconocimiento y servir de base cuando existen ambigüedades como por ejemplo en ambientes ruidosos.

Las características dinámicas de la voz tales como la evolución de formantes forman la base del aprendizaje. Se piensa que los niños inicialmente orientan su aprendizaje a sílabas completas, reduciendo el tamaño de las unidades a lo largo del tiempo hasta llegar al nivel de pequeños segmentos de voz conteniendo evoluciones cortas, de tal forma que los niños centran el reconocimiento en evoluciones de formantes, mientras que los adultos se basan fundamentalmente en secciones cortas de voz que incluyen parte de la barra de oclusión.

En [OHD95] se examinan las diferencias existentes en el uso de diferentes características acústicas para hallar el lugar de articulación en consonantes oclusivas, centrándose en la importancia de las evoluciones de los formantes en el reconocimiento atendiendo a las edades de los oyentes. Los resultados conseguidos contradicen en parte las ideas expresadas en el párrafo anterior, en primer lugar, la evolución de los formantes no se presenta como una característica espectral obligatoria en el reconocimiento realizado por niños, por otra parte, la información de corta duración es empleada no solo por los adultos, sino también por los más jóvenes.

Se confirma el emplazamiento de la información fundamental, centrada en un intervalo entre 15 y 25 ms. con 10 ms. empleados en el comienzo de la vocal posterior a la consonante.

La barra de explosión, a pesar de contribuir a que la señal de voz no posea características estacionarias, en las consonantes oclusivas [SMI94], nos ofrece de forma aislada información interesante que puede ser analizada.

Entre los estudios de fonética acústica existentes, algunos se basan en la determinación de la importancia que las barras de explosión tienen en el reconocimiento de las consonantes oclusivas. En [BON96] se investiga sobre la percepción que aportan las barras de explosión y la ayuda que juegan las vocales posteriores, para ello se realizan varios experimentos basados en verificar la habilidad de diversos oyentes para identificar secuencias con la barra de explosión aislada y en otros casos unidas a vocales posteriores.

La principal conclusión que se obtiene en este estudio es que las porciones de voz correspondientes a barras de oclusión aisladas (sin contener ningún segmento vocálico), aportan información muy fiable acerca del lugar de articulación de las consonantes oclusivas. El porcentaje de aciertos en la identificación se sitúa en el 87%, sin embargo, no se puede obtener una perfecta identificación del lugar de articulación sin recurrir al resto de factores (tamaño completo de la barra de explosión y entorno de formantes vocálicos) presentados simultáneamente.

Otros autores [BLU79] con resultados similares, concluyeron que las barras de explosión de forma aislada proporcionan pistas invariantes e independientes del contexto suficientes para la identificación de las consonantes oclusivas, sin embargo, la posición y evolución de los formantes de la siguiente vocal son necesarios para el perfecto reconocimiento de las secuencias de voz.

En [RAN95] se propone un método de identificación de las consonantes explosivas sordas, este grupo presenta características de señal de naturaleza no estacionaria, con lo que la dificultad para su correcta clasificación aumenta. El método propuesto se centra en la identificación de características de la barra de explosión utilizando análisis espectral básico. Aunque los resultados (porcentajes de acierto) no son brillantes, el método propuesto proporciona una vía para afinar los porcentajes conseguidos.

Fuera de la información aportada por las barras de explosión, el objetivo del trabajo realizado en [MUJ90] es la medición de la duración de las transiciones en la secuencia "oclusiva sorda+vocal" del castellano. Aunque diversos trabajos muestran la importancia de la evolución de los formantes en la determinación del punto de articulación consonántico, ésta investigación se centra

únicamente en la duración de las transiciones de la vocal posterior a las consonantes oclusivas sordas del castellano. La concreción con que se ha fijado el objetivo de este trabajo, ayuda a conseguir resultados más fiables que en otras investigaciones de carácter general.

El método utilizado en el estudio es el siguiente:

1.- Grabación de las 15 sílabas que se pueden obtener combinando las consonantes oclusivas sordas con las vocales castellanas colocadas en posición posterior.

2.-Selección del área de transición (cambios bruscos de frecuencia situados junto a la barra de explosión).

3.- Selección de la vocal completa (transición+área estable).

4.- Medición de tiempos.

5.- Experimentos de percepción auditiva de los grupos seleccionados.

Además de los valores concretos de tiempos obtenidos, como aportaciones del trabajo se obtiene la siguiente conclusión: la duración de las transiciones desde el punto de vista perceptivo por sí sola no es suficiente para identificar el punto de articulación de las consonantes oclusivas sordas.

Un objetivo muy complejo y no tan estudiado como otros en el área del tratamiento automático del habla, se basa en la capacidad para extraer y aislar conversaciones que se producen simultáneamente. En el campo de la percepción de la voz, se debe tener en cuenta la capacidad de los oyentes para comprender el habla en presencia de diferentes sonidos, incluidas diferentes conversaciones simultáneas. Estudios recientes han mostrado las pistas que permiten a los hablantes separar los sonidos de dos conversaciones. Los dos principales factores encontrados son: las diferencias en la frecuencia fundamental (F0) y el patrón de continuidad de formantes, esto es, la tendencia que las resonancias del tracto vocal tienen a variar de forma lenta y continua sus frecuencias a lo largo del tiempo.

En [ASS95b] se realiza un estudio encaminado a determinar cómo los oyentes explotan la característica de continuidad de formantes a lo largo del tiempo, con el fin de separar los sonidos creados por dos hablantes diferentes, para ello, se elaboran varios experimentos en los que se sintetizan sonidos, se presentan a distintos hablantes y se establecen conclusiones con los resultados obtenidos.

En el estudio mencionado, los oyentes fueron capaces de identificar vocales emitidas en paralelo más eficientemente cuando una de ellas iba precedida o seguida por transiciones de formantes

forzadas por consonantes líquidas o por glides, aunque solamente con pequeñas mejoras. Una explicación a estos resultados podría venir dada por recientes estudios que apuntan hacia la 'estrategia de cancelación'. De acuerdo con esta teoría, uno de los sonidos que compiten por el reconocimiento es eliminado o 'cancelado' de la mezcla de dos voces mediante la identificación de algún atributo que distinga la voz que interfiere, por ejemplo su estructura de armónicos o timbre. Los resultados de [ASS95b] podrían ser interpretados como un ejemplo del proceso de cancelación.

Un año más tarde, el mismo autor publica nuevos estudios relacionados [ASS96] que amplían las conclusiones obtenidas. En esta ocasión, se presentan dos modelos que tratan de predecir los efectos combinados de la transición de formantes y posición de la frecuencia fundamental con el objetivo de la identificación de vocales pronunciadas simultáneamente.

Los dos modelos predicen con suficiente claridad que las transiciones de los formantes (al comienzo o final de las vocales) generalmente no ayudan a los oyentes a identificar la vocal a la que se asocian estos formantes, por el contrario, se identifican con mayor seguridad las vocales aisladas. Las transiciones de los formantes son beneficiosas simplemente porque proporcionan al oyente una pista que ayuda al reconocimiento del habla en las regiones de transición entre vocales y consonantes.

Las siguientes referencias han sido escogidas para ofrecer una muestra del abanico de posibilidades de investigación que se puede tomar en los aspectos de la fonética acústica menos conocidos. Cabe resaltar la importancia que los estudios clásicos de identificación de formantes, tono de voz, etc. tienen en el desarrollo de todas las áreas del tratamiento automático de la voz.

En [KUS95] se realiza un experimento en el que se hacen grabaciones de 9 personas pertenecientes a 2 familias diferentes, al presentarse estas grabaciones para su reconocimiento se encontró un número significativo de aciertos en los casos en los que hablante y oyente coinciden en una misma familia. Estos resultados como cabría esperar se hacen más patentes a medida que la duración de las grabaciones aumenta hacia el tamaño de la oración, donde el factor de entonación prosódico actúa con una mayor intensidad.

Un sencillo experimento nos podría mostrar la capacidad humana para detectar variedades dialectales, y esta capacidad de reconocimiento se mantendría en unos parámetros razonables a pesar de las diferentes cualidades oratorias de los hablantes. En [TRE95] se diseña un estudio en

el que se realizan grabaciones de frases y de segmentos /hVd/. Los hablantes escogidos pertenecen a dos grupos: uno de raza blanca y el otro de personas de color. Los resultados nos muestran un porcentaje de aciertos muy significativo que mejora en las frases y disminuye en los segmentos escogidos.

[WAN95] ofrece un estudio en el que se pretende detectar emociones atendiendo a diversos parámetros físicos, para ello se escogió a ocho actores que grabaron frases en las que se expresan distintos sentimientos (ira, alegría, estado normal, nerviosismo, odio, etc.), los resultados conseguidos muestran la existencia de dos tipos de emociones: las 'no ambiguas' y las 'ambiguas', estas últimas (p.e. tristeza y depresión) se caracterizan por presentar valores paramétricos parecidos, por el contrario, en las expresiones en las que no existe ambigüedad, se consigue un buen porcentaje de clasificación.

En [COU95] se compara la capacidad para realizar identificación de vocales y discriminación de formantes de varios jóvenes sin problemas auditivos y de ancianos con disminuciones leves de percepción auditiva. Se presentaron 4 vocales a diferentes niveles sonoros (70 y 95 db.). En el caso de los ancianos la media de discriminación se situó en 69% y 80% respectivamente (con una fuerte varianza). Los jóvenes acertaron en una proporción muy cercana al 100%. Los resultados son muy similares para el primer formante, sin embargo, la tasa de aciertos en la discriminación de F2 se acercó significativamente entre los dos grupos, lo que sugiere que la identificación vocálica se consigue en parte gracias a la capacidad de los individuos para discriminar diferencias espectrales en las regiones en las que se sitúa el segundo formante.

Un correcto aprendizaje de la producción del habla requiere de la capacidad de escuchar ejemplos como mecanismo de aprendizaje/corrección [MON83]; para ello, existe la necesidad de llevar a cabo complejas sesiones de corrección de la pronunciación a las personas con deficiencias auditivas; en estos periodos de entrenamiento se utilizan logopedas altamente cualificados, cuyas habilidades sería conveniente simular en sistemas automáticos.

En [CHE95] se estudian las características acústicas que conforman una de las anormalidades más comunes que contribuyen a reducir la comprensión del habla producida por personas sordas: la nasalización en las vocales.

La corrección de la nasalización se complica debido a que es un fenómeno que suele ir unido a otro tipo de problemas en la pronunciación, por otra parte, la corrección humana provoca

decisiones subjetivas que pueden confundir a los alumnos, finalmente, no es posible realizar una inspección visual del tracto nasal para corregir estos problemas desde un punto de vista articulatorio.

Con el fin de detectar las diferencias existentes entre hablantes con algún nivel de sordera y aquellos que oyen con normalidad, en primer lugar se realizaron grabaciones (entre miembros de los dos grupos) escogidas por su contenido nasal, después se obtuvieron los espectros de voz y se compararon entre sí. Del estudio de los espectros se deduce la aparición de un pico adicional en las vocales nasales, que en el caso de las personas sin sordera se sitúa alrededor de los 950 Hz. y en el grupo con minusvalía sobre los 930 Hz. Cuanto mayor sea la amplitud de este pico mejor será la nasalización conseguida.

La principal aportación del estudio se basa en la determinación del factor A1-P1 como medida de cuantificación de la nasalización (A1 $\Rightarrow$ amplitud del primer formante, P1 $\Rightarrow$ amplitud del pico extra).

La capacidad de aproximar las características de la señal de voz de un hablante a otro, presenta una serie de aplicaciones entre las que se encuentran: normalización del habla para su posterior reconocimiento, síntesis de voz adaptada a un patrón concreto, avances en identificación del hablante, etc.

En [SLI95] se presenta un trabajo de modificación del hablante basado en la variación de los polos obtenidos con técnicas LPC, desplazando estos polos, se consigue cambiar las posiciones de resonancia del tracto vocal (que determinan los formantes).

Las etapas realizadas son:

1.- Grabación de palabras de referencia obtenidas de varios hablantes.

2.- Segmentación y etiquetado manual de las palabras en porciones fonéticas significativas.

3.- Obtención de polos mediante LPC.

4.- Aproximación de los polos de un hablante hacia los de otro, modificando los ángulos y radios de los ceros mediante transformaciones lineales simples basadas en parámetros estadísticos (medias y varianzas).

Los resultados obtenidos evidencian el mal funcionamiento de este método ante situaciones de coarticulación, por lo que se deduce la necesidad de utilizar funciones más complejas (no lineales) en la etapa de aproximación de los polos de un hablante a otro.

El objetivo fundamental perseguido en [BOO96] es determinar como el reconocimiento de sonidos del habla empeora reduciendo sucesivamente los detalles espectrales. Este estudio pretende ayudar a comprender y cuantificar el impacto que la baja resolución y la introducción de ruido en el habla produce en la percepción de la voz para personas con problemas de audición.

El método seguido se basa en la aplicación de señales ruidosas de distintos anchos de banda sobre porciones de voz. Los límites de la comprensión aparecieron al aplicar filtros de 250 Hz y 8000 Hz. El reconocimiento por parte de los oyentes de palabras aisladas se presenta más susceptible a la introducción de ruido que la comprensión de fonemas aislados.

En [LEE96] se realiza un estudio sobre los efectos producidos por la introducción de una señal ruidosa en el habla y por la incorporación de un filtro que suaviza los picos y los valles espectrales. Una vez realizadas pruebas auditivas con diferentes señales degradadas se constata el hecho de que las personas sin problemas auditivos presentan una mayor tolerancia a este tipo de efectos.

En [ROS96] se investiga en el área del modelado de la voz con y sin presencia de un sistema automático que presente parámetros en tiempo real. En este caso, aunque el aprendizaje se realiza sobre nociones de canto, los resultados se podrían extrapolar a la enseñanza asistida del habla. Las conclusiones del estudio apuntan hacia una influencia directa y positiva en el aprendizaje al utilizarse los medios automáticos de visualización de información.

En [ZAH93] se realiza un estudio muy interesante que nos alerta sobre el peligro de despreciar información espectral importante al restringirnos únicamente a las características habituales como los formantes y el tono de voz.

La búsqueda de características invariantes de los sonidos permanece como uno de los principales y más complicados problemas en el campo del reconocimiento del habla. En el caso de las vocales, desde la publicación del artículo de Peterson y Barney [PET52] se han tomado los tres primeros formantes como la fuente fundamental de información espectral.

Cuando nos enfrentamos al reconocimiento del habla continua, debemos recurrir a un conjunto de información adicional como la evolución de los formantes a lo largo del tiempo [LIN67], cuya detección requiere técnicas más complicadas que un análisis estático de polos [SMI94], [BRO89]. Actualmente se utilizan conjuntamente datos espectrales estáticos y dinámicos (representando evoluciones a lo largo del tiempo) [GOT80], [STR89].

Restringirse a una representación espectral de formantes plantea una serie de consideraciones, tales como la crítica a la reducción de información que se realiza y a la inexacta localización que en algunos casos se produce en la posición de los formantes [BLA82].

En el trabajo realizado por Zahorian y Jagharghi [ZAH93], se realiza un experimento de clasificación automática de vocales, usado para comparar las cualidades del método clásico de extracción de formantes como característica espectral frente al propuesto en el estudio, que se basa en la utilización de una envolvente suavizada del espectro completo. Los resultados básicos obtenidos apuntan hacia un mejor comportamiento (en este caso mejor clasificación) de la envolvente suavizada del espectro frente a los formantes aislados, sin olvidar que en el método clásico se maneja menos información, con las ventajas que esto conlleva.

En cuanto a las pruebas auditivas realizadas, el método propuesto se presenta más adecuado, lo cual no es sorprendente, puesto que se alberga más información en esta representación espectral que en el caso de tomar solamente formantes aislados. Una característica importante, es la mayor facilidad que existe para determinar trayectorias de picos (cimas) a lo largo del tiempo basándose en la envolvente espectral, frente a los algoritmos más complejos necesarios en el caso de los formantes [SCH95] [PLA95].

Con estas referencias se finaliza el capítulo dedicado al estudio del arte, de cuya lectura se puede deducir que aunque el campo tratado es muy amplio, toda evolución en los conocimientos, métodos y herramientas destinados a la caracterización de los formantes del habla, repercutirá positivamente en los desarrollos que se realicen en casi todas las áreas del tratamiento automático de la voz.

4 CARACTERIZACIÓN ESPECTRAL DE SONIDOS

4.1 INTRODUCCIÓN

En este capítulo se detallan los algoritmos ideados para obtener características espectrales relevantes del habla, centrándose en la determinación de los formantes que existen en la voz. El punto de partida del estudio es el método de predicción lineal (LPC), puesto que la hipótesis en la que nos basamos centra los desarrollos en este método matemático.

El planteamiento inicial es trabajar sobre las funciones espectrales básicas obtenidas mediante LPC, determinando diferentes opciones para detectar y realzar los polos. Esta información es fundamental para localizar las posiciones de los formantes y deducir sus evoluciones.

Dado al carácter empírico con que se enfoca la investigación, unido a la naturaleza del problema, las soluciones son de tipo no lineal y algorítmicas. La documentación de las metodologías ideadas, así como las etapas que conducen a su determinación, se presentan a lo largo de este capítulo con la ayuda de numerosas gráficas de funciones espectrales en dos y tres dimensiones. Aunque para caracterizar los sonidos sordos es muy importante determinar la evolución de los formantes de los sonidos sonoros adyacentes, también resulta adecuado idear métodos que actúen sobre las funciones espectrales de las consonantes implicadas, por ello, se dedica un apartado del capítulo para detallar las soluciones propuestas a este problema.

Cada uno de los subapartados del capítulo se centra en un tipo específico de sonidos. Las soluciones encontradas se adecuan específicamente a cada uno de estos tipos, y los resultados se comentan separadamente en cada sección de conclusiones. El siguiente capítulo, dedicado a complementar los estudios existentes en fonética acústica española, tomará como única herramienta de análisis los resultados que aquí se proporcionan.

4.2 EXTRACCIÓN DE FORMANTES EN SEGMENTOS VOCÁLICOS

4.2.1 RESUMEN

En el siguiente apartado se presenta un método para la estimación automática de los tres primeros formantes en un intervalo de tiempo correspondiente a un sonido vocálico. Este método resulta suficientemente genérico como para poder ser aplicado a diferentes sonidos sonoros no vocálicos.

El punto de partida para la obtención de los formantes más representativos son los parámetros PARCOR obtenidos mediante predicción lineal (LPC).

A diferencia de algunos métodos tradicionales que buscan picos en el círculo unidad en LPC o directamente en los resultados de una transformada de Fourier, en este caso se elegirán "ciertos" máximos de la función derivada segunda de la curva obtenida al evaluar la función de transferencia en el interior del círculo complejo.

Los resultados obtenidos son lo suficientemente robustos como para conseguir una buena extracción final de formantes incluso con los algoritmos más simples de suavizado de señal.

Se incluyen resultados en forma de espectrograma y comparaciones con métodos tradicionales de obtención de formantes.

4.2.2 INTRODUCCIÓN

El sonido del habla puede ser modelado como la respuesta del tracto vocal a una serie de pulsos. Las frecuencias de resonancia se manifiestan en el espectro con energía máxima. Se les denomina formantes y constituyen una información de vital importancia en el reconocimiento del lenguaje hablado [QUI93].

Según se establezcan el punto y modo de articulación de los órganos bucales, obtendremos diferentes valores de frecuencias en los formantes. Cuanto mayor sea el abocinamiento de la boca, menor será la frecuencia del segundo formante (F2) [MAR94]. Cuanto mayor sea la

apertura bucal, mayor será la frecuencia del primer formante (F1). La figura 4.1 muestra un ejemplo de la posición de los formantes F1, F2 y F3 usando un hablante masculino.

i	e	a	o	u	
267	489	711	489	267	F1
2112	1889	1222	889	711	F2
2840	2490	2540	2400	2250	F3

Figura 4.1. Posiciones medias de los 3 primeros formantes de las vocales españolas.

En la figura 4.2 se muestran las posiciones de los 3 primeros formantes del espectro de las vocales "i,e,a,o,u", pronunciadas por el autor. Los formantes de esta figura han sido obtenidos mediante el método que se explica en este apartado.

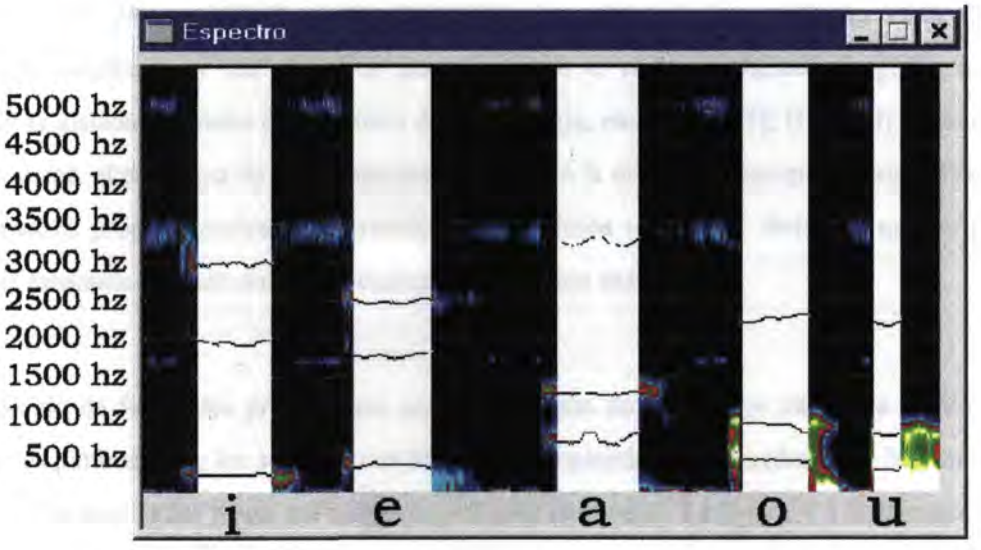


Figura 4.2. Posiciones de los formantes F1, F2 y F3 de las vocales españolas en un hablante tomado como ejemplo.

Como se puede observar, los formantes de este hablante se sitúan aproximadamente de la siguiente manera:

i: F1 ⇒ 250 Hz F2 ⇒ 1995 Hz F3 ⇒ 2950 Hz
e: F1 ⇒ 450 Hz F2 ⇒ 1800 Hz F3 ⇒ 2490 Hz
a: F1 ⇒ 700 Hz F2 ⇒ 1250 Hz F3 ⇒ 3250 Hz
o: F1 ⇒ 450 Hz F2 ⇒ 900 Hz F3 ⇒ 2300 Hz
u: F1 ⇒ 260 Hz F2 ⇒ 750 Hz F3 ⇒ 2200 Hz

Las posiciones obtenidas se asemejan mucho a las esperadas en la figura 4.1, no obstante, en otros ejemplos las diferencias pueden ser mayores debido a la variabilidad en el habla según el sexo, edad, etc.

Los formantes F1 y F2 determinan la naturaleza de los sonidos vocálicos, mientras que F3 resulta útil para discriminar sonidos consonánticos, estudiando la evolución del formante en las vocales adyacentes.

La determinación de los formantes es fundamental en la modelización del habla, y por lo tanto, la obtención de un método automático de cómputo que los calcule, resulta de gran interés en los campos de síntesis y reconocimiento de voz [PIC95], [MAR90b].

En muchos casos, el reconocimiento automático del habla ha sido abordado mediante el uso de técnicas de aprendizaje paramétrico, normalmente cadenas de Markov o redes neuronales [FRE93], [TOH92], [RAB89]. Los parámetros utilizados son usualmente los coeficientes LPC o los valores de una FFT [RAB93], [ROW92]. La calidad de los resultados varía según las técnicas empleadas y los objetivos deseados (uno o varios hablantes, lenguaje conexo o palabras aisladas, tamaño del conjunto de aprendizaje, etc.) [CAS87], [CAS90], pero en todos estos casos, el problema de fondo planteado radica en la distancia conceptual existente entre los parámetros proporcionados y los sonidos que se desea reconocer, distancia que se pretende cubrir mediante los métodos de aprendizaje automático existentes.

El cálculo de formantes proporciona una información de gran valor situada a medio camino entre los parámetros y los sonidos, con lo que la complejidad de las cadenas de Markov o redes neuronales empleadas puede ser reducida de forma sustancial. La figura 4.3 ilustra esta idea.

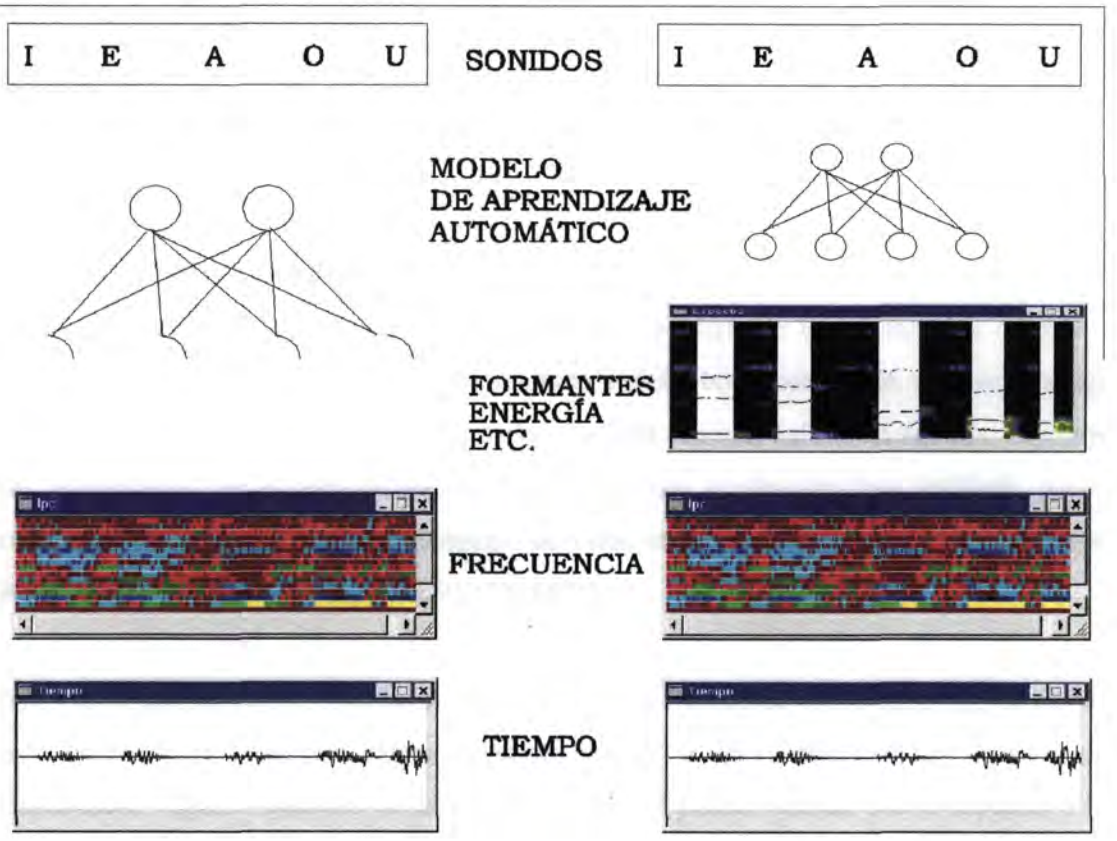


Figura 4.3a

Figura 4.3b

Niveles usuales empleados en el reconocimiento del habla, cargando la importancia del proceso en el método de reconocimiento a), o en el cálculo de características de nivel medio b).

Desde un punto de vista computacional, la opción b) de la figura 4.3, requiere una etapa complementaria, aunque por otra parte, los modelos de aprendizaje automático se reducen. En cualquier caso, el objetivo es conseguir resultados (convergencia) en la etapa de aprendizaje, demasiado complicada en el caso a) de la figura 4.4, así como poder emplear los formantes como modelo del habla, más cercano a nosotros que los parámetros matemáticos de bajo nivel.

Los siguientes apartados se estructurarán de la siguiente manera:

- Introducción al análisis de predicción lineal
- Enfoques para abordar el problema
- Entorno de trabajo y desarrollo
- Conceptos básicos en los que se basa el método
- Detalles del algoritmo usando dos niveles de abstracción
- Esquema del método

- Suavizado posterior de la señal
- Conclusiones

4.2.3 PREDICCIÓN LINEAL

Puesto que el método de obtención de formantes ideado parte de los coeficientes obtenidos mediante predicción lineal, el siguiente apartado repasa brevemente este tipo de análisis matemático desarrollado en el capítulo anterior. Se hará especial énfasis en aquellos aspectos más relacionados con el resto de los apartados. Para una explicación más detallada, puede consultarse literatura clásica de tratamiento de señal orientada al procesamiento de la voz [ROW92], [PAR86], [RAB78], [RAB93], [MAK75].

El problema que plantea LPC radica en obtener un cierto número de coeficientes (k), los cuales serán usados para conseguir alguna característica de la señal mediante una función de transferencia. Algunos ejemplos de características que se pueden obtener son el espectro o la predicción de datos adicionales a la señal (Figura 4.4).

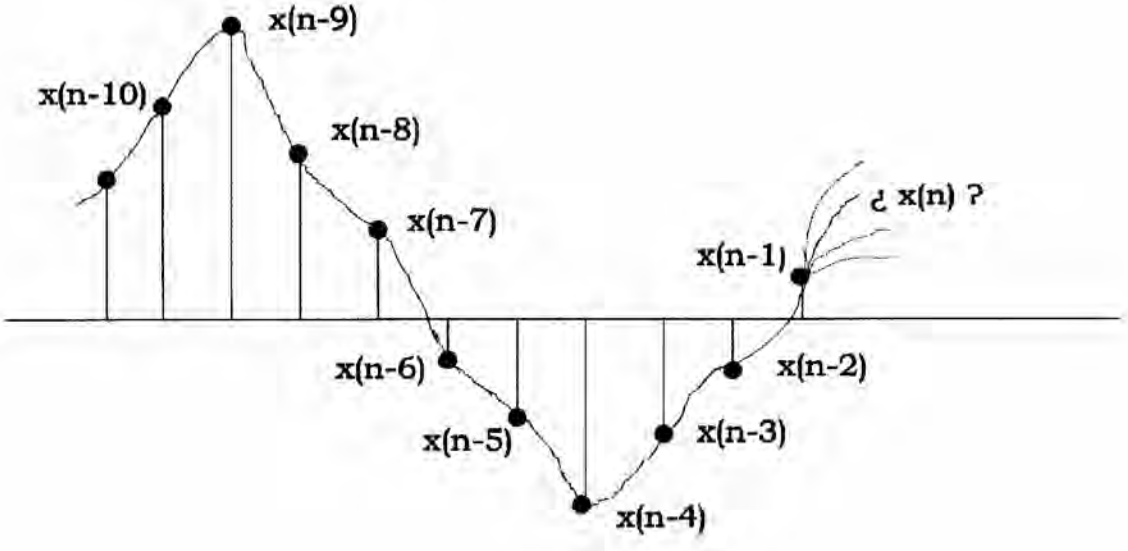


Figura 4.4

Predicción de los datos adicionales de la señal aplicando LPC

$$e(n) = x(n) - s(n) = x(n) - \sum_{k=1}^{p} a_k(n)\, x(n-k) \qquad (1)$$

El objetivo es dar un valor, *s(n)*, lo más aproximado al valor que se desconoce, *x(n)*, es decir, una predicción del valor *x(n)* cuyo error sea el menor posible. Esto implica que LPC no va a ser un método exacto, sino que tan sólo dará valores aproximados, aunque esto hará que se decremente el tiempo de cómputo frente a la transformada rápida de Fourier.

Los valores a_k desconocidos, se calculan minimizando el error *e(n)*, para lo cual se aplican mínimos cuadrados. Para ello se forma el error cuadrático medio en el intervalo '*n*' que se desea considerar.

$$L = \sum_{n} e^2(n) \qquad N\text{-}1 \geq n \geq 0 \qquad (2)$$

Igualando $\partial L / \partial a_j$ a 0 para j=1,2,....p y simplificando, se obtiene:

$$\sum_{k} a_k \, \varnothing_{jk} = x_j \qquad j=1,2,...p \qquad (3)$$

$$k=1,2...p$$

donde:

$$\varnothing_{jk} = \sum_{n} x(n-j)\, x(n-k) \;,\; x_j = \varnothing_{j0}$$

Una vez que se han obtenido los coeficientes a_k, para hallar el espectrograma, se evalúa la magnitud de la función de transferencia:

$$H(z) = \frac{a_0}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (4)$$

donde:

$$z = r\, e^{j(2\Pi n/N)} \qquad n = 0,1,.... N\text{-}1$$

Aumentando el valor de *N* se consigue más resolución frecuencial, a costa de un aumento lineal del cómputo necesario.

Haciendo *r=1*, la solución se obtiene en el borde del círculo unidad, mientras que con valores de *r<1*, los resultados se hallan en el interior del círculo, con búsqueda en la circunferencia de radio r. Esta posibilidad de selección en la búsqueda de soluciones ha sido muy empleada en la determinación de formantes muy cercanos, cuyos picos aparecen unidos en el círculo unidad y

han de ser resueltos con valores de $r<1$ (situación típica en vocales nasalizadas) [SCH70], [CAN74].

La figura 4.5 representa un instante de tiempo de señal de voz, cuyo espectro ha sido evaluado con valores de $r \in (0.8..1)$. El instante captado corresponde al sonido 'o' de la figura 4.2. El eje X representa la frecuencia de 0 a 5500 Hz, con una definición de N=128 puntos. En el eje Y se observan los valores obtenidos tras aplicar la función de transferencia a los parámetros LPC. El eje Z contiene 11 funciones espectrales calculadas en distintos radios (r) del círculo unidad.

Dada la función de transferencia implementada, los ceros de la señal se han hecho coincidir con los picos de la figura. Para $R=1$ (circulo unidad) los 3 primeros picos corresponden a los formantes F1, F2 y F3. Desgraciadamente en la realidad la obtención de los formantes no es tan inmediata como en este caso.

La figura 4.5 ilustra como a medida que disminuimos el radio en la función de transferencia, los picos de la curva resultante se hacen menos intensos hasta el punto de desaparecer. Esta es la razón de que usualmente se busquen los formantes con radio $r=1$, y sólo en el caso de existir problemas se descienda por el círculo complejo.

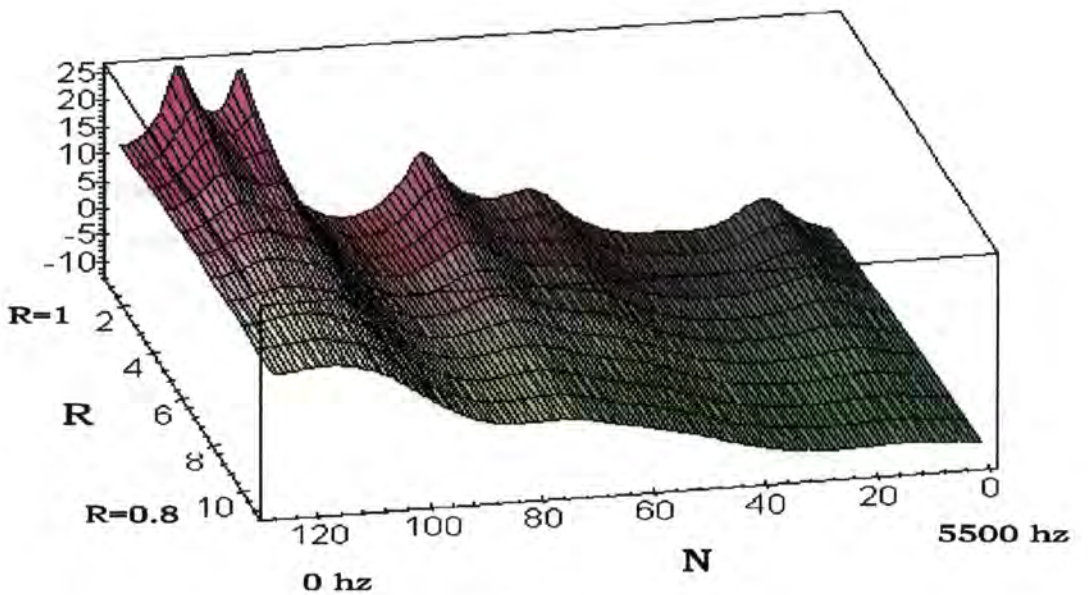


Figura 4.5

Evolución de las funciones espectrales obtenidas aplicando diferentes radios a la función de transferencia

4.2.4 ESTRATEGIAS PARA LA OBTENCIÓN DE FORMANTES

Matemáticamente, se pueden obtener los formantes de un instante de sonido hallando los polos del filtro, para ello basta con calcular las raíces de las ecuaciones planteadas y seleccionar aquellas soluciones que se adapten a los formantes esperados utilizando criterios algorítmicos. Este método presenta la ventaja de ser conceptualmente muy simple, pero existe el grave inconveniente de que requiere una gran carga de computación para ser llevado a cabo.

Otra posibilidad con tiempos de computación mucho más bajos consiste en el cálculo de la función espectral evaluada usualmente en el círculo unidad. Con dicha función, se seleccionan los máximos y se escogen (también con criterios algorítmicos) los 3 picos que se consideren más adecuados para representar a F1, F2 y F3 [SCH70].

La determinación de picos sobre una curva espectral presenta varios problemas de exactitud:

1. No siempre se encuentran todas las soluciones en el radio evaluado sobre el círculo complejo.
2. A veces aparecen de forma temporal pequeños picos que no deben ser confundidos con formantes.
3. Cuando dos polos están muy cercanos, tienden a unirse en un solo pico, pudiéndose perder de esta manera un posible formante.

La figura 4.6 ilustra el caso de un pico que representa un formante, pudiéndose apreciar su evolución a lo largo de 6 instantes de tiempo en mitad de una grabación de la vocal 'o'. En $t=1$ la pendiente de subida al pico no es lo suficientemente grande como para poder etiquetarlo como candidato a formante, a medida que pasa el tiempo, el máximo se va consolidando, hasta que se convierte en el candidato a F2 de la vocal. Esta es la situación opuesta al caso 2 descrito anteriormente.

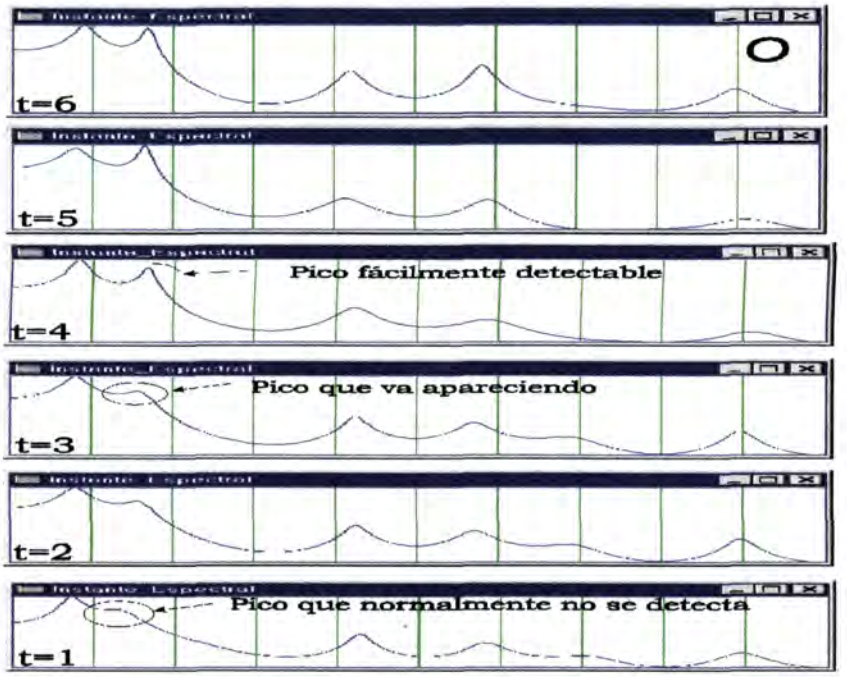


Figura 4.6

Ejemplo de evolución a lo largo del tiempo de los picos en las funciones espectrales

El método de obtención de formantes que se explica en este apartado no se basa en la resolución de ecuaciones en busca de polos, con lo que aunque los tiempos de ejecución son adecuados, se debe asumir un grado de inexactitud en los resultados que tal y como ocurre en la naturaleza, se pueden intentar corregir en niveles superiores del proceso de reconocimiento, por ejemplo realizando comprobaciones de consistencia a nivel de palabra, sintáctico, etc. [CAS87]

4.2.5 DESARROLLOS PRELIMINARES

Los gráficos presentados en el artículo como ventanas Windows, pertenecen a una aplicación informática desarrollada para probar las distintas teorías que se van creando con el fin de obtener resultados satisfactorios en diversos campos del procesamiento de la voz. De esta manera se va trabajando en un ciclo de teoría $\Rightarrow$ programación $\Rightarrow$ prueba $\Rightarrow$ error , que va generando la suficiente documentación y unidades orientadas a objetos como para poder imaginar las siguientes teorías y probarlas hasta obtener los aciertos esperados.

Se ha implementado también un método clásico de obtención de formantes partiendo de LPC y evaluando la función de transferencia en el círculo unidad, seleccionando picos como

candidatos y aplicando finalmente técnicas de suavizado de señal. Este método ha sido utilizado como punto de partida para evaluar los resultados de los demás, como 'mejoras' que se obtienen frente a los formantes que se calculan en esta técnica 'tradicional'.

4.2.6 CONCEPTOS BÁSICOS DEL MÉTODO

Obtener formantes a partir de los picos proporcionados por una FFT, tiene el inconveniente de que la función espectral presenta demasiada información, por lo cual resulta difícil elegir los picos que representan los formantes principales, descartando el resto de los máximos. Al usarse LPC, trabajamos con una función simplificada, que a modo de envolvente de la FFT conserva la información fundamental, disminuyéndose aquella que no es necesaria para nuestro propósito [RAB93].

Aplicando un razonamiento similar al del párrafo anterior, podemos usar LPC variando el radio 'r' de aplicación de la función de transferencia, con lo que obtendremos diferentes niveles de detalle en la información que se usará en la determinación de formantes.

La figura 4.7, al igual que la figura 4.5, presenta la evolución de las funciones espectrales entre $r=0.8$ y $r=1$ (en el eje Z) de un instante de tiempo perteneciente a la vocal 'o'. Como se puede apreciar, los polos de la señal, tienden a presentarse cercanos a $r=1$ (círculo unidad), mientras que según nos alejamos de este valor, la información que se muestra va perdiendo rápidamente nivel de detalle.

El método de obtención de formantes propuesto en este artículo, se basa en la extracción de la función espectral a partir de un $r<1$ tal que se mantenga la información fundamental, eliminando detalles superfluos. En la figura 4.7, se puede observar como la función en $r=1$ presenta un nivel de detalle excesivo, especialmente (en este ejemplo) en frecuencias bajas, donde entre los dos primeros picos representantes de F1 y F2 de la 'o' existe un pequeño máximo local no representativo. Además se presentan mínimos demasiado pronunciados y poca uniformidad en los valores alcanzados por los máximos (eje Y).

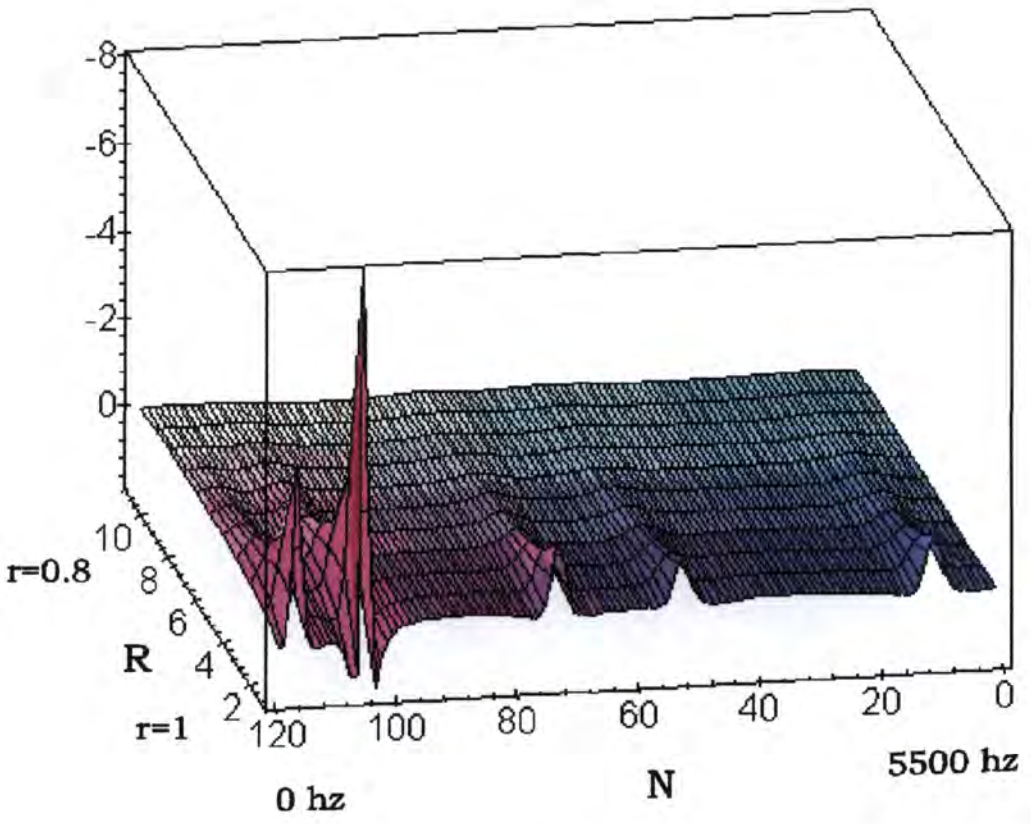


Figura 4.7

Ejemplo de evolución de las funciones espectrales entre r=0.8 y r=1 en un instante de tiempo de la vocal 'o'

A la vista de la figura 4.7, se puede apreciar que resultaría más adecuado escoger un $r<1$, pero no demasiado alejado del círculo unidad para no perder información representativa. La figura 4.8 es equivalente a la 4.7, pero con un intervalo de $r \in (0.85, 0.95)$, lo que elimina los valores excesivos de la señal en $r=1$ y focaliza el gráfico en el área del círculo complejo en el que pretendemos escoger el valor de *'r'*.

Observando la figura 4.8, podemos determinar que existen diversos valores razonables de *'r'* que podemos escoger. Idealmente deberíamos optar por aquel que minimice el error en la función objetivo (formantes de la señal), como obviamente no disponemos de una función matemática que poder minimizar, escogeremos un valor representativo, en nuestro caso $r=0.9$.

Una buena aproximación a este problema es la visión 'topológica' del mismo, según la cual, el espacio de soluciones se traduce en picos que representan los polos y depresiones que significan

mayores valores de la función de error, tal y como aparece en las figuras tridimensionales presentadas. En este caso, en lugar de 'caminar' por la cuerda de la montaña uniendo cimas, caminamos por la 'ladera' de valor $r=0.9$, notando bajo nuestros pies la pendiente lateral que en su evolución llegará a los picos cercanos a $r=1$.

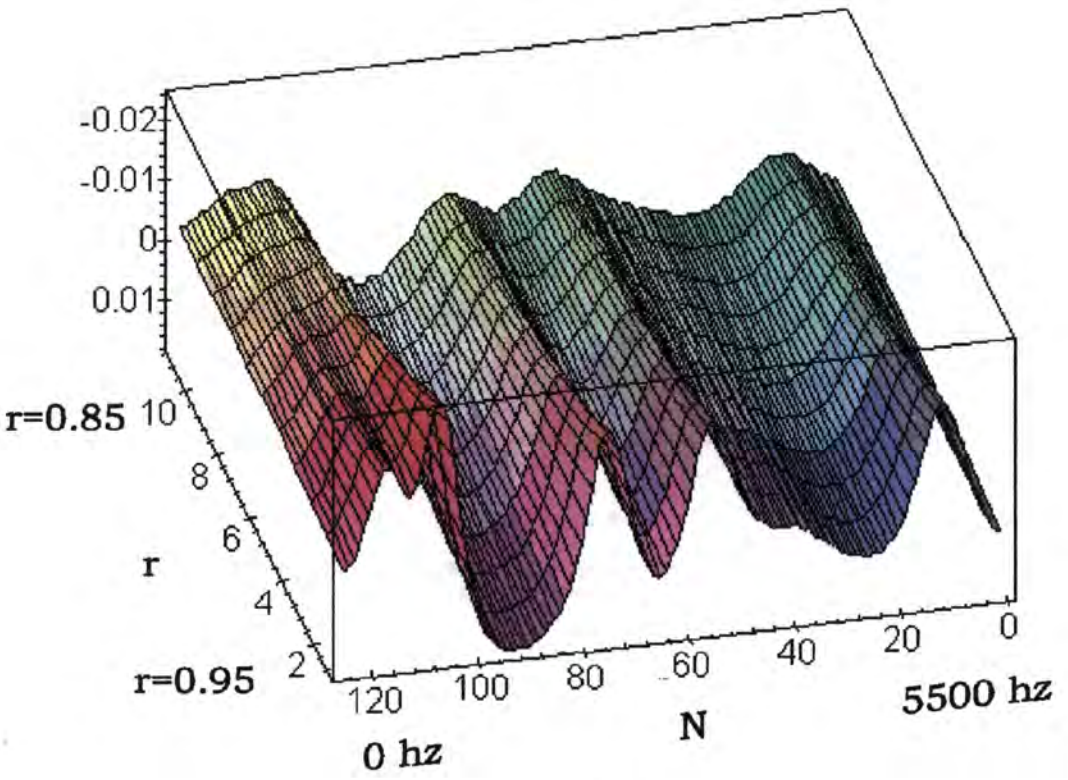


Figura 4.8
Ejemplo de evolución de las funciones espectrales entre r=0.85 y r=0.95 en un instante de tiempo de la vocal 'o'

Si asumimos la premisa razonable de que en la señal de voz los polos no aparecen de forma instantánea, sino que se forman gradualmente, las 'montañas' que nos encontramos en $r=0.9$, representarán las zonas de información más representativa, de tal manera que en sonidos vocálicos una 1ª montaña en bajas frecuencias determinará una 'o' o una 'u', una 1ª montaña en frecuencias más altas una 'i' o una 'e', mientras que la 'a' vendrá dada por valores intermedios. La figura 4.9 presenta 5 gráficos de ejemplo correspondientes a la función espectral en $r=0.9$ de las vocales españolas.

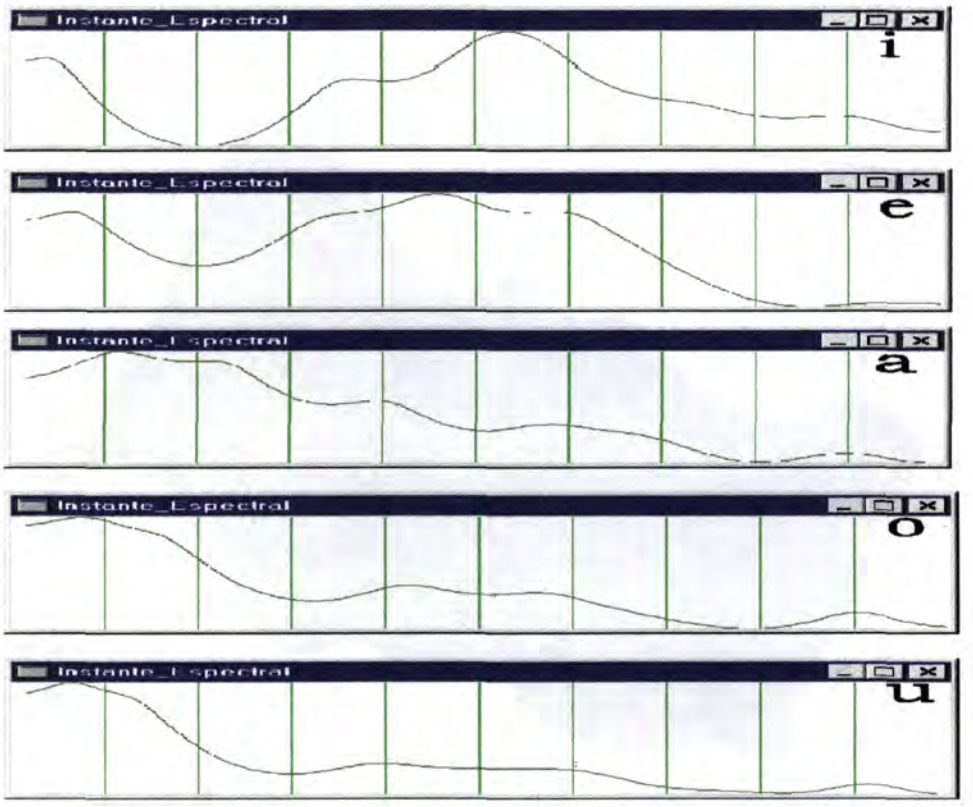


Figura 4.9

Ejemplo de función espectral evaluada en r=0.9 de las 5 vocales españolas

Con la estrategia explicada, resulta más fácil determinar el tipo de sonido vocálico que investigando los detalles de la función espectral en *r=1*. Sin embargo para hallar los formantes exactos de la vocal, aparentemente deberíamos realizar la búsqueda en *r=1*. La figura 4.10 presenta las funciones espectrales de un intervalo de tiempo (eje Z) de la señal de voz de una vocal 'o'. Como se puede observar, los mayores valores de la función se presentan en frecuencias bajas ('o', 'u'), pero el detalle de la situación de F1 y F2 se pierde.

Avanzando en nuestro ejemplo del montañero, nos surge la pregunta. ¿ Es posible distinguir los picos que se presentarán en la cima *(r=1)* caminando por la ladera *(r=0.9)*?. Aplicando esta pregunta al ejemplo presentado en la figura 4.10., ¿Sería posible distinguir los 2 polos que conforman F1 y F2 analizando de izquierda a derecha la primera 'montaña' del gráfico?. La respuesta es que sí. Se puede obtener una muy buena aproximación, pero es necesario disponer de unos 'pies' muy sensibles que distingan no sólo la pendiente del camino, sino además la variación de esta pendiente, de tal forma que variaciones bruscas de pendiente en *r=0.9* significarán cimas en *r=1*.

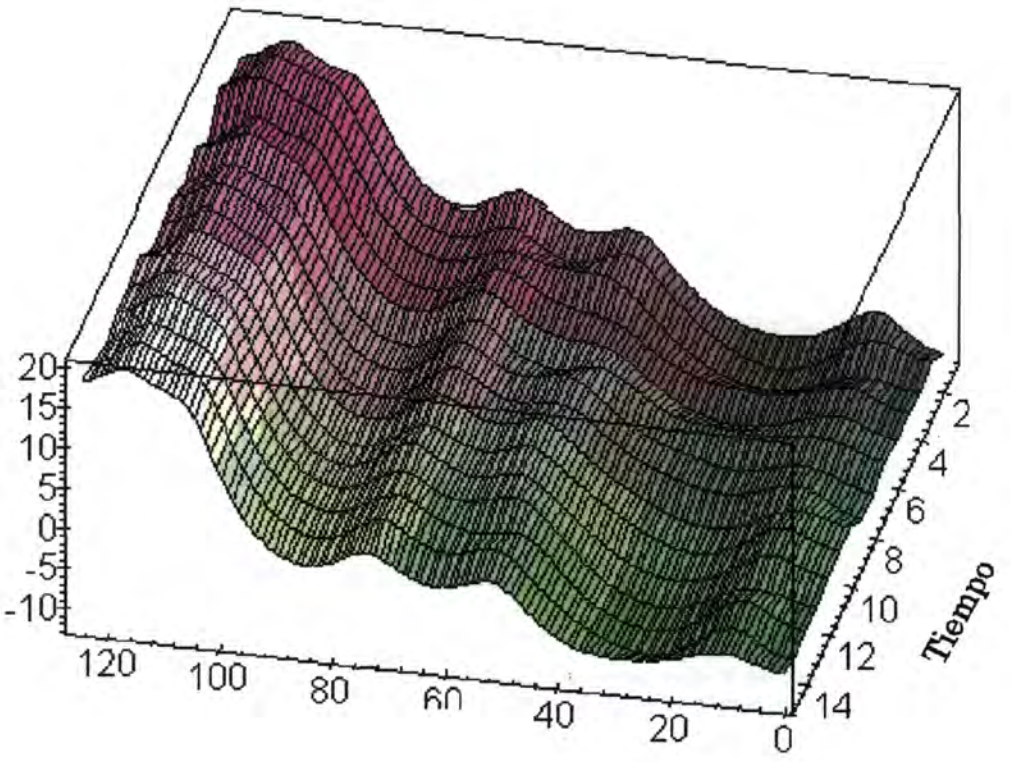


Figura 4.10

Función espectral evaluada en r=0.9 de un intervalo de tiempo de una señal de voz correspondiente a la vocal 'o'

Matemáticamente, esto significa que debemos hallar la derivada segunda de la función espectral obtenida en *r=0.9*, esperando que las variaciones de la pendiente sean suficientemente significativas como para determinar los formantes de la señal. La figura 4.11 muestra las segundas derivadas de funciones espectrales análogas a las de la figura 4.9. En este caso, cada pico representa un formante, pudiéndose diferenciar perfectamente F1 y F2, incluso de forma más simple que analizando la señal en el círculo unidad (figura 4.5).

La figura 4.11 complementa a la 4.9, añadiendo a las funciones espectrales representadas, sus derivadas segundas. La combinación de ambas funciones proporcionará la información básica que empleará el método propuesto de obtención de formantes.

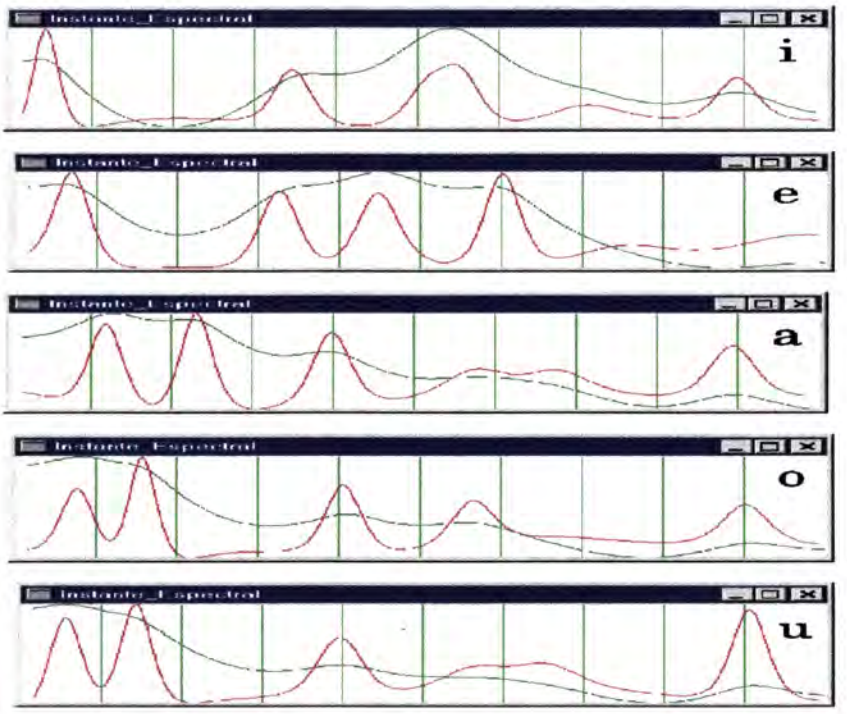


Figura 4.11

Ejemplo de función espectral evaluada en r=0.9 de las 5 vocales españolas junto a su derivada segunda

4.2.7 PRIMERA APROXIMACIÓN AL ALGORITMO

A la vista de los resultados mostrados en la figura 4.11, parece lógico determinar como formantes los máximos de la derivada segunda de la función espectral evaluada en $r=0.9$, sin embargo, aunque normalmente este método funcionaría correctamente, existen situaciones en las que no se puede asociar de forma directa picos y formantes, usualmente debido a la aparición de máximos 'extras' que deberían ser ignorados.

A continuación se muestran tres ejemplos obtenidos en tres vocales diferentes pertenecientes a distintos hablantes. Estos ejemplos se apoyan en gráficos de la derivada segunda de la señal en $r=0.9$ durante un intervalo de tiempo de aproximadamente 60 ms.

El primer caso se trata de una vocal 'i' en cuyo espectro el segundo pico no representa un formante (figura 4.12), el eje X nos proporciona la información frecuencial y el eje Z la temporal.

El segundo caso muestra una vocal 'a' en la que de los tres primeros máximos se obtienen los dos primeros formantes, por lo que tampoco aquí se puede asociar de forma biunívoca

máximos con formantes (figura 4.13). Como se puede observar, en los primeros instantes de tiempo existen sólo dos máximos que pasan a convertirse en tres en las últimas realizaciones espectrales.

El tercer y último caso de ejemplo, presenta el espectro de la señal evaluado con función de transferencia aplicada en $r=0.9$ y la derivada segunda de esta señal. La vocal analizada es la 'e' y como se puede observar, el primer máximo se convierte en dos a medida que avanza el tiempo, con lo que resulta difícil determinar la situación de los formantes (figura 4.14).

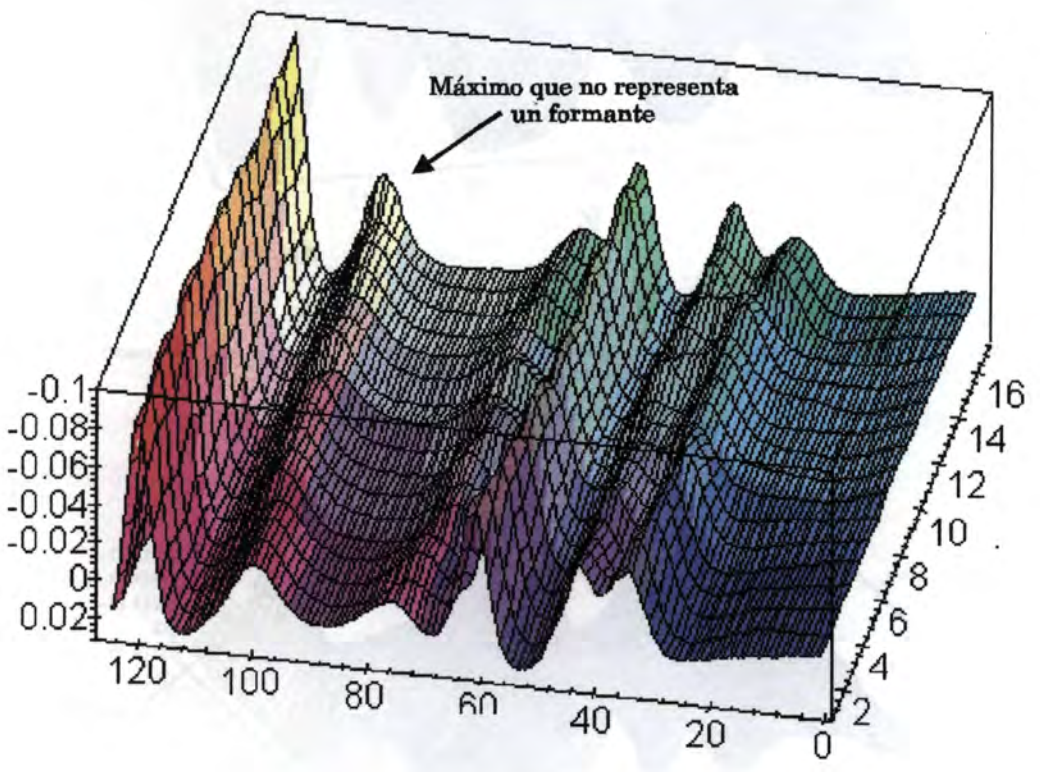


Figura 4.12

Caso de una vocal 'i' en la que aparece un máximo que no corresponde con un formante.

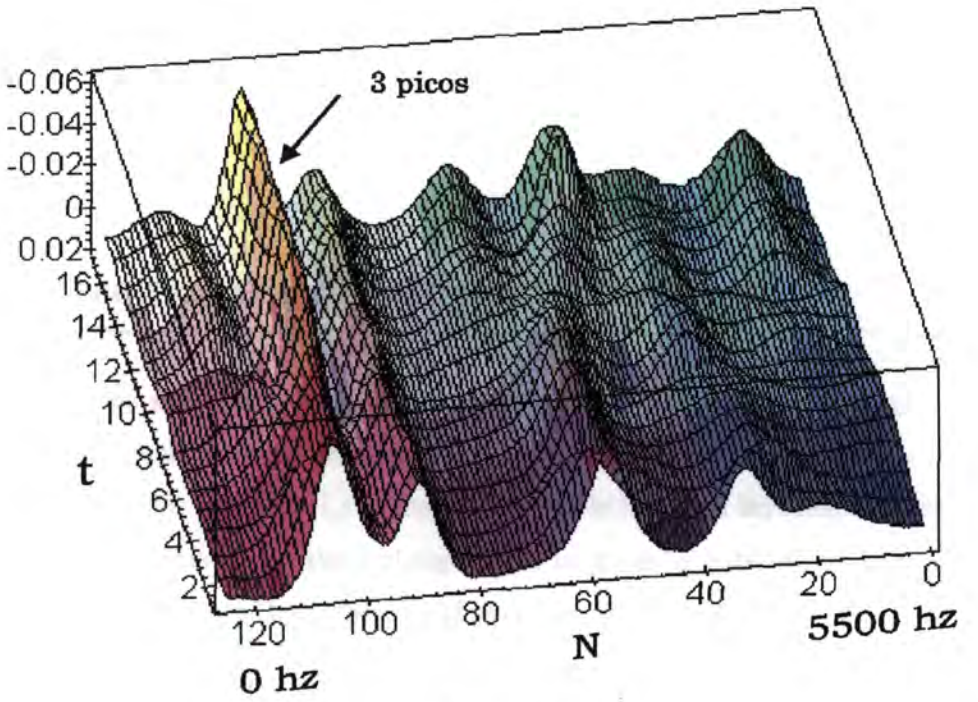


Figura 4.13

Caso de una vocal 'a' en la que los máximos pasan de representar formantes de forma directa a no hacerlo.

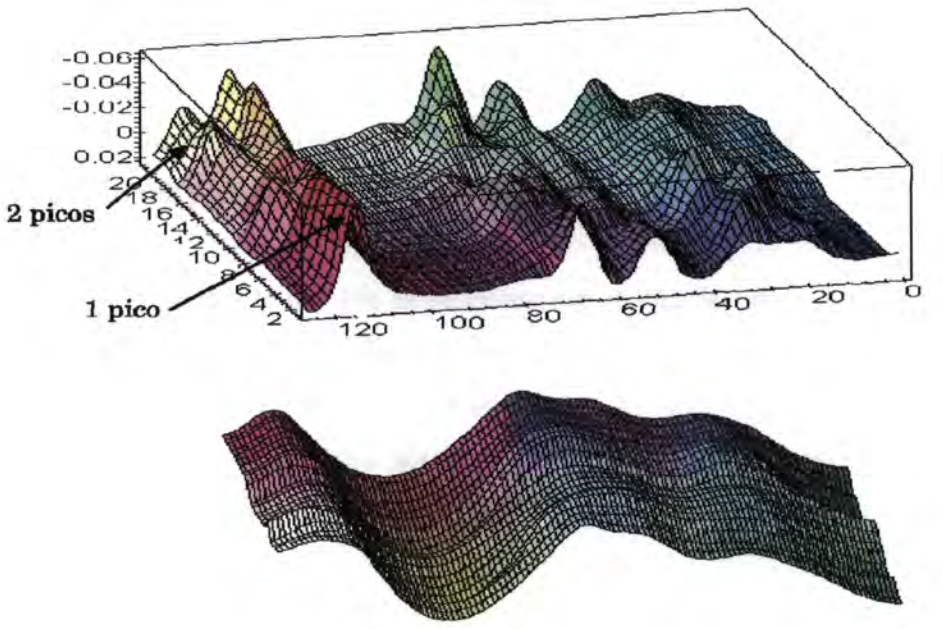


Figura 4.14

Caso de una vocal 'e' en la que resulta difícil determinar sus formantes al duplicarse un máximo en bajas frecuencias.

Tanto las situaciones que muestran los ejemplos anteriores como otros casos similares que se producen, aún siendo excepciones en las señales de voz, se dan con una frecuencia suficiente como para catalogar de inadecuado el sencillo algoritmo de extracción de formantes esbozado en este apartado. No obstante la solución final se basará en el método propuesto, ampliándolo y mejorándolo mediante el análisis de la función base de la cual se extrae la derivada.

4.2.8 SEGUNDA APROXIMACIÓN AL ALGORITMO

Tal y como se explicó en el apartado anterior, basándose en el ejemplo que proporciona la figura 4.9, trataremos de determinar la posición de un mínimo global que divida la función espectral en dos secciones. La primera sección (a la izquierda del mínimo) albergará el primer formante de todas las vocales y el segundo de la 'a', la 'o' y la 'u'. La segunda sección (a la derecha del mínimo global) albergará el tercer formante de todas las vocales y el segundo de la 'i' y la 'e'.

El mínimo global deberá ser buscado en frecuencias menores a aproximadamente 4 KHz, con el fin de evitar los valores pequeños de señal que habitualmente se presentan en frecuencias altas del espectro, en las que por otra parte no se encuentra ninguno de los 3 primeros formantes de la voz en español.

Ocasionalmente, en el caso de la vocal 'a', el mínimo global de la función espectral no divide la curva en las dos secciones esperadas, en esta situación el algoritmo se puede simplificar escogiendo los tres primeros máximos como formantes.

La figura 4.15 muestra la división en secciones que resulta de aplicar a las funciones de la figura 4.11 un algoritmo de búsqueda del mínimo global entre 50 y 3400 Hz con la restricción de que la pendiente de la curva en las proximidades del mínimo sea de al menos el 4%. El mínimo se representará en adelante por un círculo en la función espectral. Los círculos en la derivada segunda indican máximos detectados en dicha función.

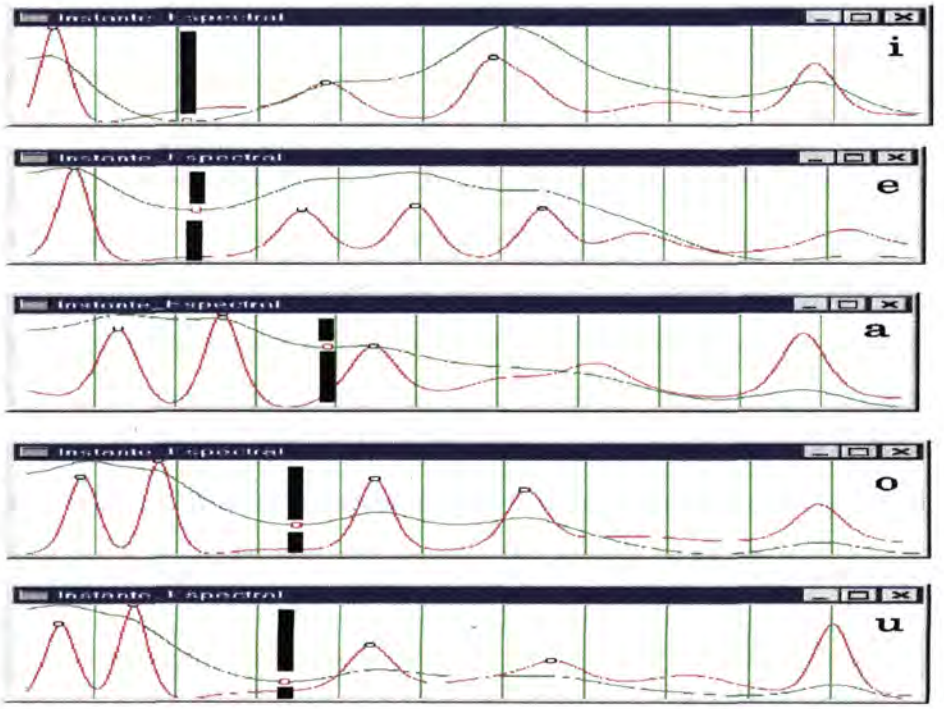


Figura 4.15

División en secciones que se obtiene al hallar el mínimo global de la función espectral.

En esta ocasión, la división del espectro se realiza de la forma esperada y los formantes se obtienen tal y como se había previsto. En la segunda sección, en la que normalmente aparecen varios máximos en la derivada segunda, siempre se escogerán de izquierda a derecha, es decir, los de menor frecuencia.

El ejemplo mostrado, aunque sirve para explicar el funcionamiento básico del algoritmo, no presenta situaciones complejas que justifiquen la utilización de un método como el que aquí se expone. De hecho, a la vista de la figura 4.11 se podría establecer que para hallar los tres primeros formantes, basta con asociarlos a los tres primeros máximos (con un mínimo de pendiente) de la derivada segunda. Por desgracia, a menudo se presentan situaciones mucho más difíciles de solucionar que la que aquí se analiza, por ejemplo las mostradas gráficamente en las figuras 4.12 á 4.14.

En algunos casos, la función espectral presenta máximos que no deberían ser asociados a formantes, sino simplemente ignorados. La figura 4.12 muestra un ejemplo típico de esta situación. Para dar una solución a este tipo de problemas, en el método propuesto se calcula el valor medio de la señal espectral en cada una de las secciones halladas, y se asocian con formantes únicamente aquellos máximos (de la función derivada segunda) en los que el valor de

la función espectral en la frecuencia marcada por el máximo, es mayor que la media de la señal espectral en la sección correspondiente.

La figura 4.16 presenta cuatro casos en los que el procedimiento propuesto ignora máximos que no deberían ser tomados en cuenta, con lo que la situación es tratada correctamente. Los dos primeros gráficos se corresponden con dos instantes de tiempo de la figura 4.12. Los 2 segmentos de recta horizontales representan el valor de la media de la señal y tamaño de cada una de las secciones en las que dividimos la función espectral. Como se puede observar, en la letra 'i', el segundo máximo de la derivada 2ª se encuentra en una frecuencia en la que el valor de la media de la función espectral (recta horizontal) es mayor que el valor de la señal espectral evaluada en $r=0.9$, por lo que el pico es ignorado (no presenta círculo en la cima).

En los gráficos correspondientes a una 'o' y una 'u' de la figura 4.16, es el tercer pico de la función derivada segunda el que se ignora. En los cinco casos se ha conseguido evitar el tomar como formante un máximo con la energía y pendiente suficientes como para que se hubiera elegido erróneamente como representativo en un método más simple de extracción de formantes.

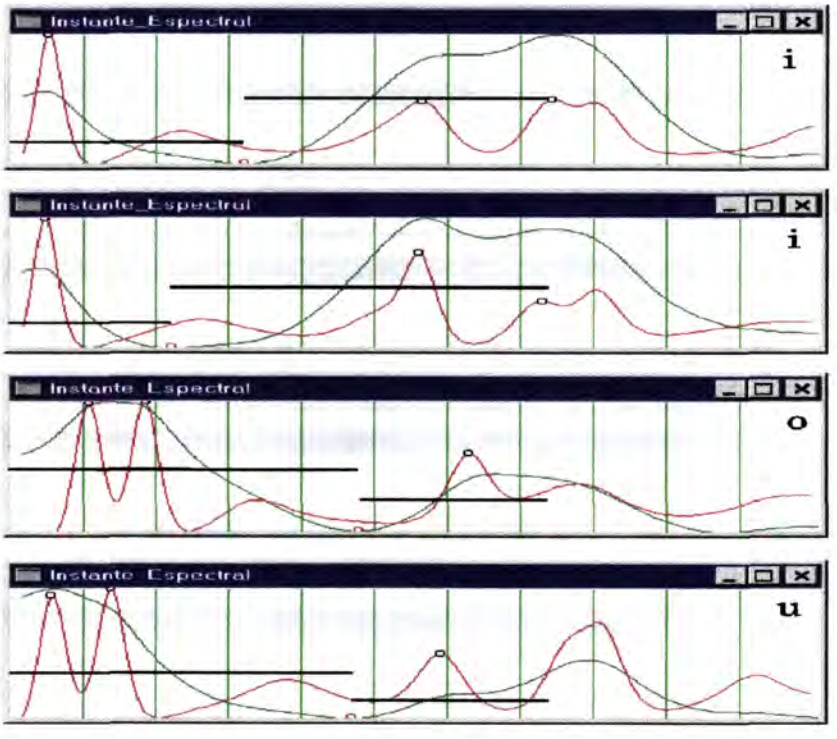


Figura 4.16

Ejemplo en el que se ignoran máximos de la función derivada segunda que no deben ser tomados como formantes.

En realidad, lo que estamos haciendo es desestimar aquellos máximos que se encuentran por debajo de la media de los valores de la función espectral, independientemente de la importancia que tenga la variación de la pendiente de los mismos.

El inconveniente de aplicar esta condición en la selección de máximos, reside en el hecho de que esporádicamente se pueden eliminar picos representativos. Para resolver esta situación basta con rebajar el valor de la media de la señal tanto como sea necesario hasta encontrar los tres formantes que buscamos.

Los tres primeros gráficos de la figura 4.17 presentan casos en los que nos encontramos la situación descrita y la resolvemos bajando el valor de la media (segmentos horizontales) tantas veces como sea necesario hasta encontrar los formantes deseados.

En el último gráfico de la figura 4.17 se representa un detalle del método propuesto, que se basa en una mejora en la determinación del mínimo global que divide el espectro en dos secciones. La mejora consiste en tomar como nuevo mínimo global el primer mínimo que se encuentra a la derecha del último máximo de la primera sección espectral.

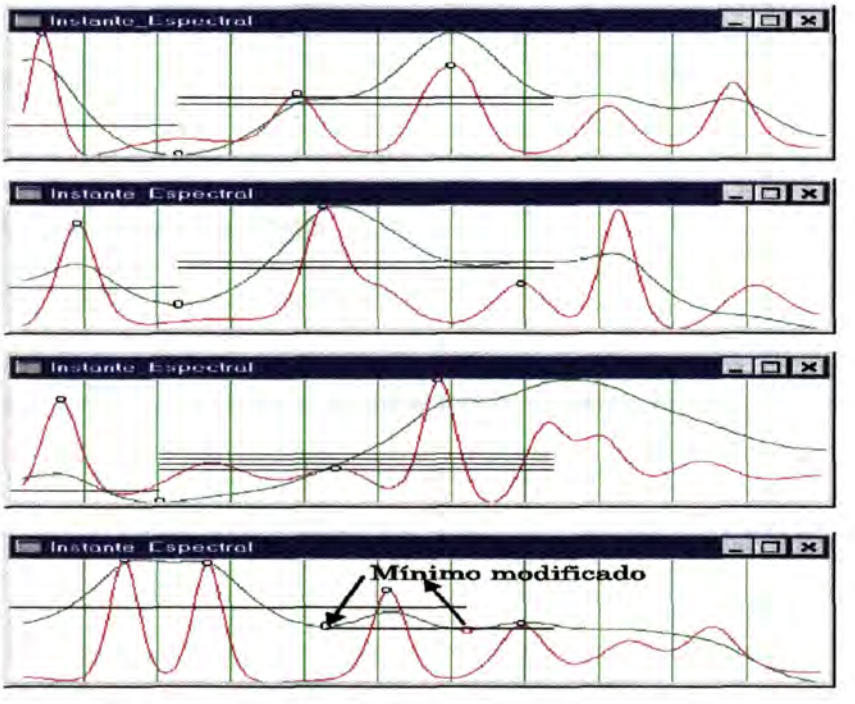


Figura 4.17

Máximos tomados como formantes después de rebajar el valor de la media de la función espectral.

Normalmente el nuevo mínimo coincidirá con el global hallado previamente, aunque en algunas ocasiones se encontrará en una frecuencia más baja y determinará un tamaño de la segunda sección más apropiado.

Por último, cuando se nos presente el caso de haber obtenido tres formantes en la primera sección del espectro (esta situación se puede dar esporádicamente en la vocal 'a' de algunos hablantes), se puede eliminar el primer pico o bien fusionar los dos primeros máximos promediando sus frecuencias espectrales. Estas acciones resuelven situaciones como la representada en la figura 4.13.

4.2.9 MÉTODO DE OBTENCIÓN DE FORMANTES

Con los conceptos explicados en las dos aproximaciones al algoritmo, se puede esquematizar el método propuesto, evitando la confusión que introduciría la inclusión de detalles que ya han sido desarrollados en los dos apartados anteriores.

En esta sección se presentarán de forma ordenada, las etapas que determinan el método de obtención de formantes propuesto en este artículo:

1.- Hallar los coeficientes LPC.

2.- Aplicar la función de transferencia con valor $r=0.9$.

3.- Pasar la función a escala logarítmica.

4.- Calcular la derivada segunda de la función anterior.

5.- Encontrar el mínimo global de la función espectral en un rango de 50 a 3500 Hz. Si no ha sido encontrado, asociar los 3 primeros máximos con los formantes buscados.

6.- Hallar la media de la función espectral antes del mínimo global.

7.- Determinar las posiciones de los máximos de la función derivada segunda antes del mínimo global.

8.- Descartar los máximos en los que el valor de la función espectral es menor que el de la media de la señal en la sección estudiada (antes del mínimo global).

9.- Si se han obtenido tres máximos, promediar los dos primeros para obtener F1 y asociar el tercero con F2.

10.- Sustituir el mínimo global por el primer mínimo situado a la derecha del último máximo detectado.

11.- Hallar la media de la función espectral entre el mínimo establecido en el paso 10 y la frecuencia 3.5 KHz.

12.-Determinar las posiciones de los máximos de la función derivada segunda después del mínimo actualizado.

13.- Descartar los máximos en los que el valor de la función espectral es menor que el de la media de la señal en la sección estudiada (entre el mínimo actualizado y 3500 Hz).

14.- Si en total se han hallado menos de tres formantes, rebajar el valor de la media de la señal espectral obtenida en la etapa número 11, hasta que se obtengan los tres formantes buscados.

El algoritmo que aquí se detalla, obtiene los 3 primeros formantes de un instante de tiempo cualquiera, y se puede aplicar independientemente de los valores de la señal de voz en instantes adyacentes.

A modo de ejemplo, la figura 18 presenta los resultados obtenidos aplicando este algoritmo sobre las señales de voz de las 5 vocales españolas pronunciadas por dos hablantes diferentes. Cada punto dibujado sobre fondo blanco representa la posición frecuencial de un formante en el instante de tiempo correspondiente.

Las barras obtenidas representan la continuidad de los formantes en las vocales pronunciadas, en el primer gráfico se analizan las vocales 'i', 'e', 'a', 'o', 'u', mientras que en el segundo la secuencia es 'a', 'e', 'i', 'o', 'u'.

En las 10 vocales analizadas, los formantes se encuentran en las posiciones frecuenciales esperadas.

En algunos casos se presenta el valor de un cuarto formante F4, que puede ser utilizado como información adicional para etapas posteriores en el caso de que las realizaciones espectrales no se obtengan tan claras como las del ejemplo.

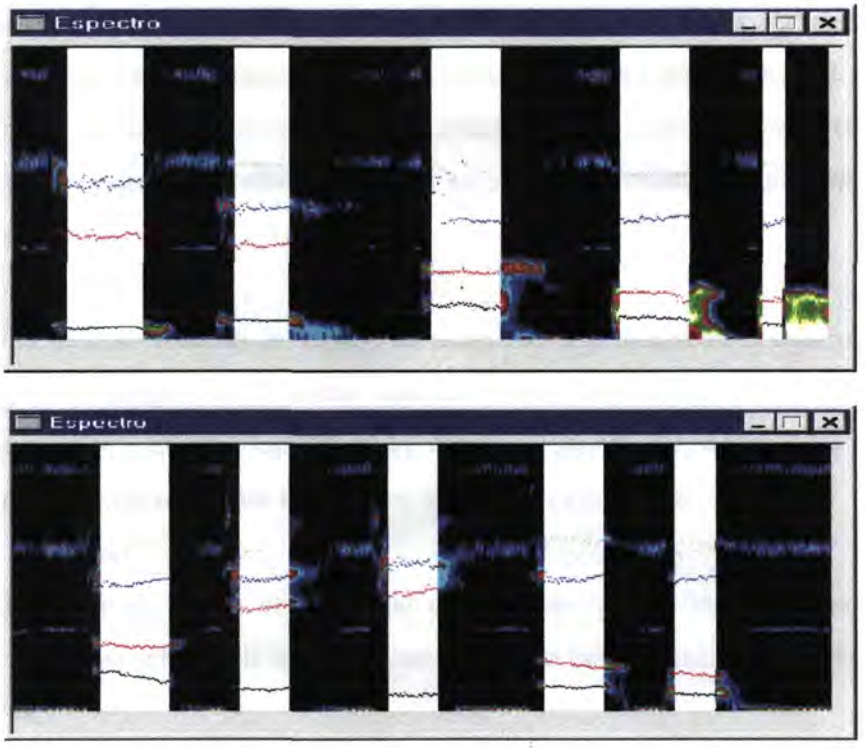


Figura 4.18

Posiciones de los formantes obtenidos en vocales pronunciadas por 2 hablantes.

4.2.10 SUAVIZADO DE LA SEÑAL

Aunque usando el método de obtención de formantes explicado en este artículo los resultados obtenidos ofrecen una buena calidad, con el fin de tratar discontinuidades, saltos bruscos o formantes erróneos, resulta adecuado añadir una etapa de suavizado de señal.

El suavizado de señal se basa en el supuesto de que en sonidos sonoros, los formantes no deberían cambiar de forma brusca a lo largo del tiempo, al igual que no lo hacen nuestros órganos articulatorios. Puesto que el método de obtención de formantes no tiene en cuenta esta característica, podemos usar la información redundante que poseemos al estudiar la señal en un intervalo temporal, de manera que se suavicen discontinuidades bruscas en la evolución de los formantes [PLA95], [SCH95], [CAN74].

En nuestro caso se han implementado dos algoritmos con distinta funcionalidad:

El primer algoritmo realiza un alineamiento de los picos seleccionados, de tal manera que si en algún caso se ha omitido un formante, los siguientes no sean confundidos. Por ejemplo, si no se ha detectado el segundo formante, el algoritmo evita que F3 sea tomado como F2. Para realizar esta función se mantienen 4 valores de medias aritméticas a lo largo del tiempo, recordando las últimas cinco muestras. Con ello se consigue asignar los formantes en un tiempo $t+1$ a los valores medios de F1 a F4 en $t-4$ a t.

El algoritmo descrito presenta la ventaja de adaptarse dinámicamente a cualquier evolución natural de los formantes y de filtrar adecuadamente los valores extremos erróneos. El inconveniente es que es muy sensible a las desapariciones de formantes. Este algoritmo ha proporcionado buenos resultados aplicado en señales sonoras de voz.

El segundo algoritmo realiza el suavizado de señal de 'grano fino', consiguiendo curvas continuas sobre una señal en la que el primer algoritmo había eliminado las discontinuidades más bruscas.

Con el objetivo de poner a prueba la calidad de los resultados en el método de obtención de formantes, se ha implementado la opción más simple de las posibles: un filtrado de señal mediante la sustitución de cada valor por el de la media aritmética de los 'n' anteriores. A pesar de la sencillez de esta etapa los resultados conseguidos han sido satisfactorios.

En el caso de querer hallar los formantes en señales sonoras de voz muy cambiantes, tales como diptongos o triptongos, basta con variar esta última etapa para obtener los resultados deseados.

4.2.11 RESULTADOS

El método de extracción de formantes ha sido probado en un gran número de señales de voz de vocales aisladas, correspondientes a distintos hablantes con diferentes edades y sexos.

En la gran mayoría de los casos se han hallado de forma correcta las frecuencias de los dos primeros formantes, aunque F3 en algunas ocasiones ha quedado difuso en una nube de puntos que no ha hecho muy fiable la utilización de la etapa de suavizado.

La figura 4.19 presenta los espectros de las 5 vocales españolas pronunciadas por ocho hablantes diferentes cogidos al azar. La figura 4.20 muestra los formantes hallados en los espectros de la figura 4.19. En estas 40 realizaciones sólo existe un error situado en el tercer formante de la vocal 'i' del último espectro (inferior derecho). Este error ha sido debido al primer algoritmo de suavizado de señal, y no a la etapa de obtención de formantes.

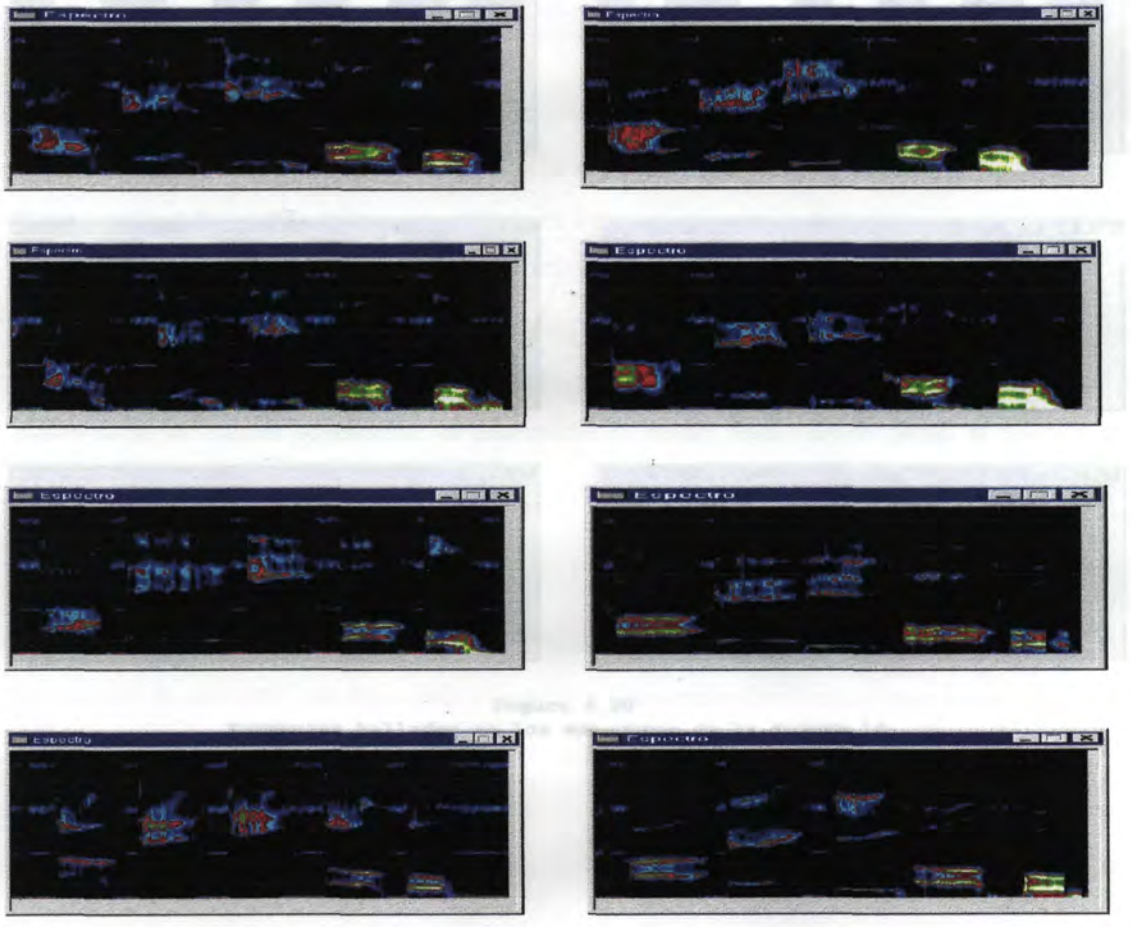


Figura 4.19

Espectros de vocales pronunciados por diferentes hablantes.

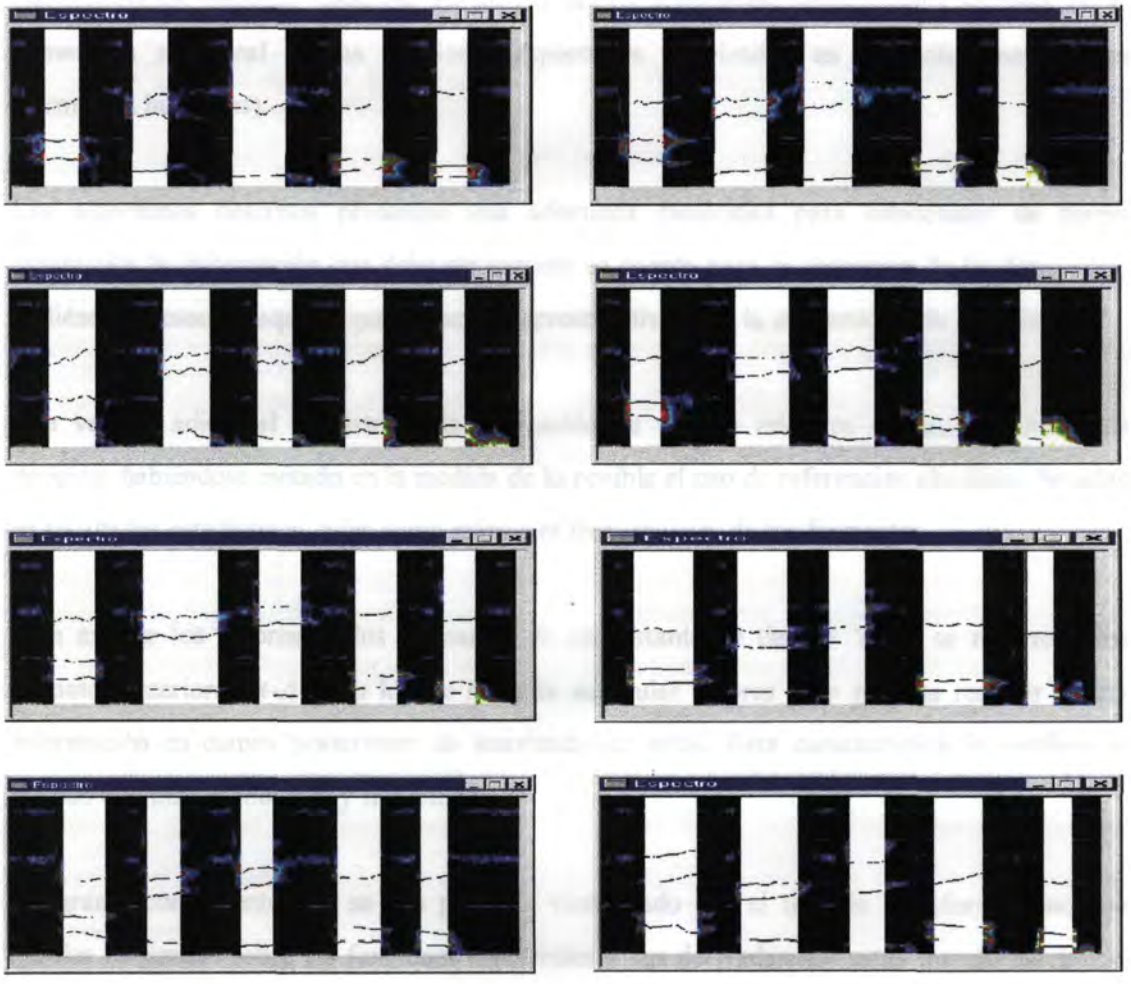


Figura 4.20
Formantes hallados en los espectros de la figura 19.

4.2.12 CONCLUSIONES

En esta sección se ha mostrado un método de obtención de formantes aplicable a sonidos sonoros. Está basado en la utilización simultánea de funciones espectrales extraídas a partir de parámetros LPC y las derivadas segundas de estas funciones.

La principal contribución de este método radica en la utilización de funciones espectrales obtenidas fuera de los valores usuales de búsqueda de polos, por lo que realmente se trabaja

con tendencias hacia la solución. El núcleo de los algoritmos desarrollados se basa en la conversión no lineal de las funciones espectrales suavizadas en las soluciones finales (formantes buscados).

Los algoritmos descritos presentan una adecuada capacidad para seleccionar de forma automática la información que debe ser tomada en cuenta para la obtención de los formantes, pudiéndose desechar aquella que es menos representativa para la consecución de este fin.

Una ventaja adicional consiste en la utilización de valores relativos en los algoritmos de decisión, habiéndose evitado en la medida de lo posible el uso de referencias absolutas basadas en resultados estadísticos, tales como márgenes frecuenciales de los formantes.

Para extraer los valores de los formantes en un instante de tiempo 't' no se recurre a los instantes anteriores 't-Δ', con lo que se evita acumular errores y se permite recurrir a esta información en etapas posteriores de suavizado de señal. Esta característica le confiere al método una mayor robustez y flexibilidad.

Un gran inconveniente que se nos presenta viene dado por el tipo de transformaciones no lineales realizadas sobre las funciones espectrales y sus derivadas. En estas transformaciones, los tiempos de cómputo son demasiado elevados y poco predecibles como para hacer posible el uso de este método en sistemas de tiempo real estricto.

4.3 DETERMINACIÓN DE CARACTERÍSTICAS ESPECTRALES EN SONIDOS SONOROS

4.3.1 RESUMEN

Partiendo de las funciones espectrales obtenidas en la sección anterior, se propone un método que mediante transformaciones matemáticas y aplicación de algoritmos, convierte la función original de obtención de formantes en una nueva, cuyas características se adecuan a la representación espectral de los sonidos sonoros de la lengua castellana.

Los algoritmos y transformaciones aplicados a las funciones base, son de naturaleza no lineal y están orientados a realzar los resultados espectrales en torno a las zonas con formantes.

Se presentan gráficas comparativas de los resultados que se van obteniendo a medida que se aplican las diversas transformaciones matemáticas, así como varios espectros finales comentados y comparados con sus equivalentes conseguidos aplicando el método clásico de predicción lineal.

4.3.2 INTRODUCCIÓN

La correcta representación del espectro de la voz, resulta de gran importancia para el desarrollo de diversos campos de la ciencia relacionados con el tratamiento, reconocimiento y síntesis del habla [MED85]. Esto es debido a que cuanto mayor sea la exactitud y claridad de los espectros, mejor será la comprensión que se logre sobre los fundamentos de la caracterización de los sonidos emitidos en la producción de voz.

La exactitud y precisión en la determinación de los parámetros espectrales de la voz incide de forma directa sobre la calidad de los procesos de síntesis y reconocimiento del habla, al proporcionar a éstos un conjunto de patrones de referencia más fiables y genéricos.

Las personas que más directamente se benefician de cualquier progreso en la representación espectral del habla son aquellas que trabajando en diversos campos de la ciencia (logopedia, filología, tratamiento automático de la voz, etc.), necesitan adquirir conocimientos de fonética acústica, y desean establecer dependencias claras entre la fonética articulatoria que establece la pauta de emisión de los sonidos y las correspondientes visualizaciones e interpretaciones en los espectros de voz.

Una vez determinada (en la sección anterior) una manera de extraer los formantes en segmentos vocálicos, para conseguir espectros de buena calidad, resulta necesario idear un método que detecte y realce los formantes en cualquier sonido sonoro.

Cubriendo adecuadamente la representación de los sonidos sonoros, no sólo se abarca la mayor parte de las realizaciones del habla, sino que además determinamos en gran medida la naturaleza del resto de los sonidos, puesto que como es sabido, la evolución de los formantes en la proximidad de un sonido sordo es en muchos casos la mejor forma para caracterizarlo [MEM78], [PIC95], [REP78].

Los sonidos sonoros básicos de la lengua española, ordenados por la abertura en el modo de articulación (sonoridad) son los siguientes:

vocales: [a], [e], [o], [i], [u] y sus realizaciones nasales: [a_{φ}], [e_{φ}], [o_{φ}], [i_{φ}], [u_{φ}]

laterales: [l,], [l.], [l], [l_{φ}], [λ]

vibrantes: [r], [rr]

nasales: [m], [μ], [n,], [n.], [n], [n_{μ}], [n_{δ}], [η]

fricativas: [β], [d.], [j], [γ]

africada: [dξ]

oclusivas: [b], [d], [g]

Los sonidos sordos que serán analizados en la siguiente sección son:

africadas: [t$\int$]

fricativas: [f], [θ], [s], [x]

oclusivas: [p], [t], [k]

El siguiente apartado tratará los métodos de obtención de las diversas funciones espectrales empleadas en la realización de los espectros de voz que se han conseguido. Después se presentarán diversos espectros en dos y tres dimensiones, comentándose las ventajas e inconvenientes de cada forma de representación, escala de colores empleada, funciones espectrales utilizadas, etc. Por último se plantearán las conclusiones más relevantes que se obtienen tras la realización de este estudio.

4.3.3 ALGORITMOS Y FUNCIONES IDEADOS

En la sección anterior se razonaba la conveniencia del uso de la función de minimización de error usando LPC con radio de búsqueda en r=0.9. A partir de esta función se hallaba su derivada segunda y se obtenía la información básica de extracción de formantes.

El objetivo que ahora se plantea es más ambicioso y complejo, puesto que hemos aumentado los sonidos a analizar de sólo los vocálicos a todos los sonoros. Por ello resulta necesario crear nuevas funciones espectrales, obteniéndolas a partir de las que ya disponemos.

La primera transformación que se ha realizado, consiste en la ponderación de la función derivada segunda en base a los valores dados por la función de minimización de error en r=0.9, con ello se minimizan los valores espectrales cuando más nos alejamos de los polos matemáticos.

La figura 4.21 muestra las dos funciones base y el resultado para diversos ejemplos vocálicos. Como se puede apreciar, en todos los casos se minimizan (aunque no siempre se eliminan) valores que podrían interpretarse erróneamente como formantes. Con la escala calibrada en 500Hz, en el primer caso 'o', casi se elimina el falso formante situado a 1800 Hz, en el segundo caso 'i', el situado en 1500 Hz, por último, en el 3° y 4° casos desaparecen por completo los formantes erróneos situados en 2000 Hz y 1500 Hz respectivamente.

El ejemplo mostrado y los siguientes se han tomado partiendo de grabaciones de voz pertenecientes a diferentes hablantes, y eligiendo como base distintas palabras de referencia.

Así mismo se debe resaltar que cada uno de los casos expuestos resulta representativo respecto a una gran variedad de tomas analizadas en detalle previamente, y que obviamente no pueden ser mostradas en su totalidad.

Los resultados obtenidos aplicando esta primera función son buenos en cuanto a su capacidad para filtrar valores espectrales erróneos; el problema reside en que también se minimizan (sin llegar a desaparecer) valores espectrales situados en el entorno de formantes correctos. Esto se puede observar muy claramente en el segundo caso de la figura 4.21, donde se aprecia la disminución de los valores correspondientes al primer formante del sonido 'i'.

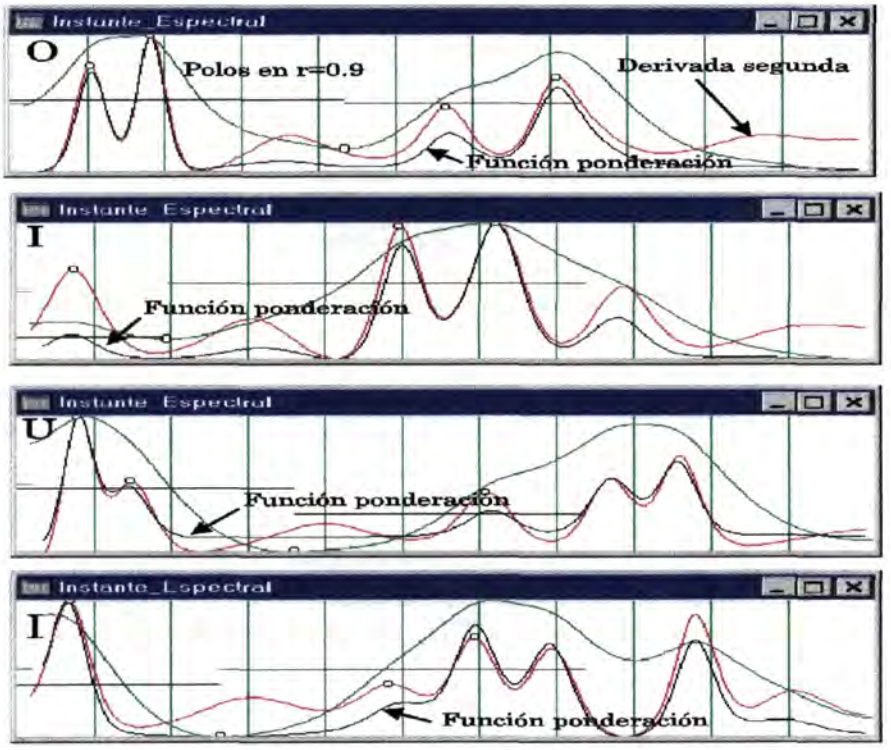


Figura 4.21

Polos, Derivada segunda y función obtenida ponderando las anteriores

En las figuras 4.21, 4.22 y 4.23, los círculos situados sobre la función derivada segunda, indican los formantes reales de los sonidos representados. Estos formantes han sido obtenidos aplicando el método explicado en la sección anterior. Las líneas horizontales se corresponden con la media aritmética de los valores de la función de minimización de error antes y después de su mínimo global, tal y como se detalló en los algoritmos de extracción de formantes en segmentos vocálicos. Estos valores nos serán de gran utilidad para la obtención de la función espectral definitiva.

Para conseguir que la función espectral no quede disminuida en las zonas donde existen formantes reales, se debe dividir la función de polos en dos partes diferenciadas por el lugar que ocupa el mínimo global de la curva. Para cada una de estas partes, se realzarán los valores en los que los polos son máximos (error mínimo) y se disminuirán donde estos valores sean más pequeños.

En la figura 4.22 se compara la función ponderación (presentada en la figura 4.21) con una nueva función en la que se realiza el realzado mencionado en el párrafo anterior.

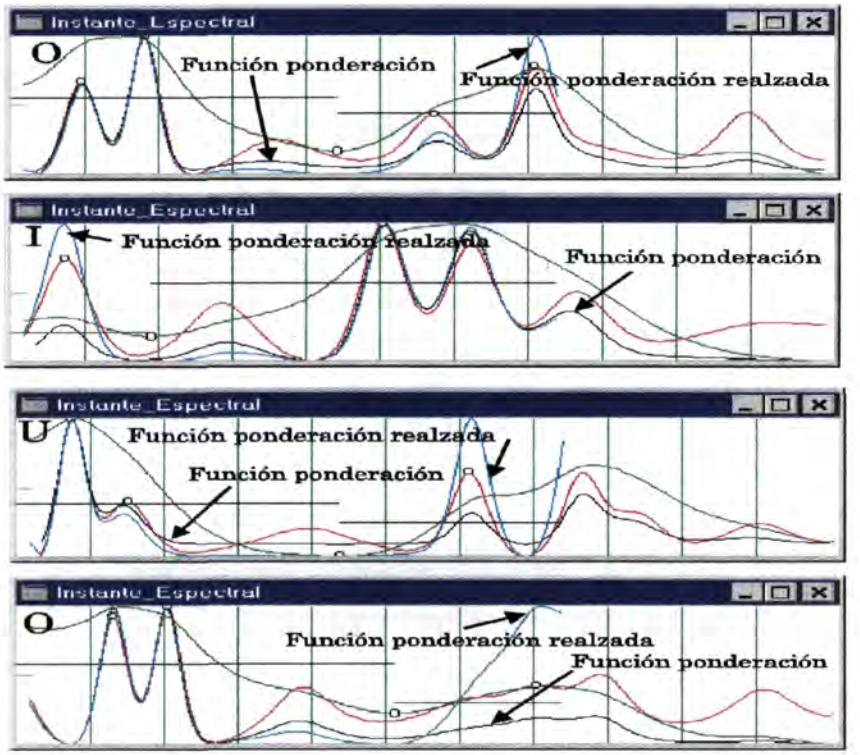


Figura 4.22

Polos, Derivada segunda, función ponderada y función ponderada realzada obtenida a partir de las anteriores

A la vista de las gráficas, se puede observar como la función obtenida realza adecuadamente los formantes. En el caso 2, perteneciente al sonido 'i', el primer formante se eleva sobre el valor de la derivada segunda, mientras que el falso formante situado en 1500 Hz se disminuye incluso más que en la función anterior.

Esta función, además de respetar y mejorar el aplanamiento de la curva en las zonas donde no existen formantes, eleva los valores en las zonas donde si los hay, con lo que resulta mucho

más adecuada que la anterior, aunque todavía existen casos donde se puede llegar a tomar como formante algo que no lo es. Esta situación se aprecia mejor en los casos 2 y 4 en las frecuencias 1500 Hz y 1850 Hz respectivamente, en donde existen unas pequeñas elevaciones que no han sido eliminadas por completo.

La función conseguida refleja fielmente las características espectrales de la voz en la práctica totalidad de las situaciones, sin embargo, a veces, se presentan ejemplos en los que la derivada segunda contiene un máximo de gran altura en una frecuencia en la que no existe formante. En estas ocasiones, la función realzada minimiza este error, pero no siempre lo hace desaparecer por completo. Esta situación se presenta en la figura 4.23 en 1800 Hz para el caso 1, 1550 Hz en el 2, 1600 Hz en el 3 y 1900 Hz en el cuarto caso.

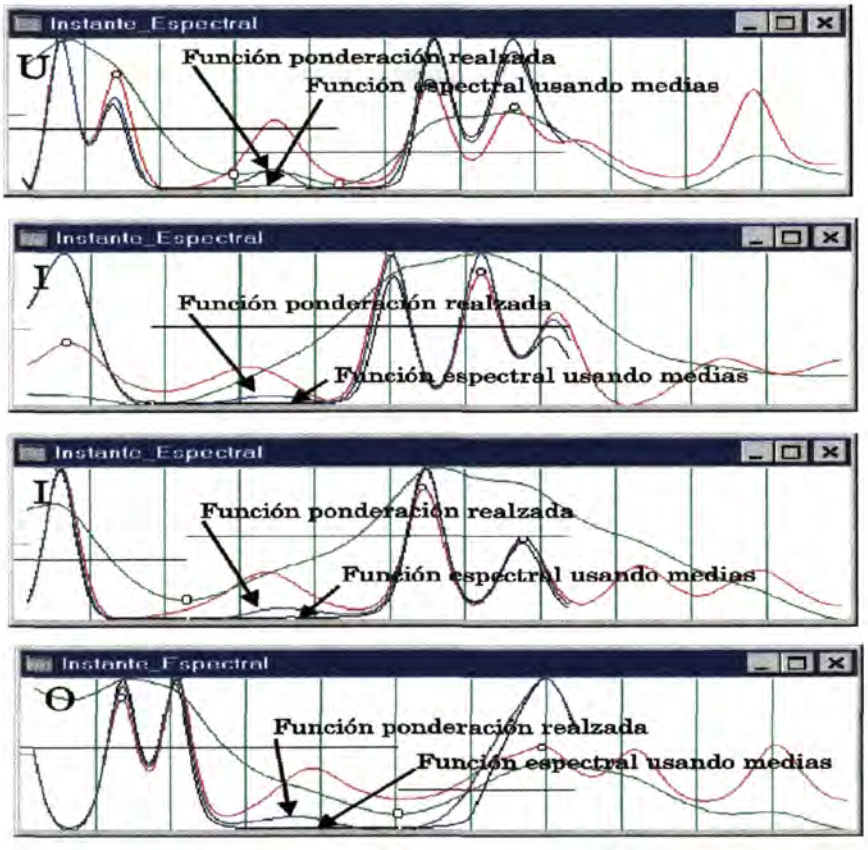


Figura 4.23

Función espectral obtenida ponderando según las medias de la función espectral de minimización de error.

Con el fin de solventar este problema, se ha ideado una función espectral en la que se realiza una ponderación entre la derivada segunda y la función de minimización de error (polos en r=0.9) basada en los valores de las medias aritméticas de ésta última (barras horizontales en las gráficas).

La función espectral objetivo se formará a partir de la derivada segunda, aumentando sus valores en las frecuencias en las que la función de error sea mayor que su media correspondiente, y disminuyéndose en las demás frecuencias. En un principio, la proporción de aumento y disminución se estableció linealmente a la distancia entre el valor de la función de error y su media asociada, pero de esta manera no se eliminan por completo las elevaciones erróneas de la función.

Finalmente se estableció una ponderación no lineal, cuyo comportamiento se ilustra en la gráfica a) de la figura 4.24, en donde el eje x representa la distancia entre la función de minimización de error y su media asociada para una frecuencia dada. Los valores positivos se aplanan logarítmicamente, para evitar crecimientos desproporcionados de los formantes. Los valores negativos se reducen mediante la inversa de la función exponencial, con lo que los falsos formantes desaparecen por completo debido a los efectos de esta función exponencial.

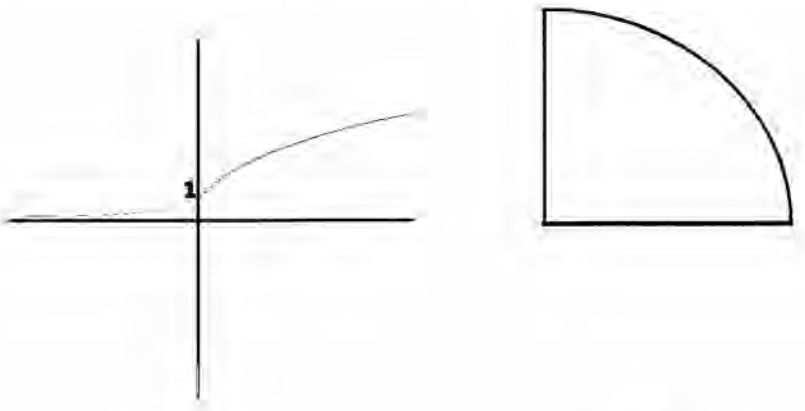


Figura 4.24

a) Función de transformación para la obtención de la función espectral ponderada según las medias ($x<0 \Rightarrow$ Inversa de la función exponencial, $x>0 \Rightarrow$ Función logarítmica).

b) Función de aplanamiento de los valores espectrales por debajo de la media aritmética de la señal de voz en tiempo.

El resultado de este conjunto de operaciones no lineales aplicados sobre las funciones base, es una función espectral cuyo comportamiento es excelente resaltando los formantes reales y haciendo desaparecer los erróneos.

Como se puede observar en la figura 4.23, la función espectral obtenida, respeta los formantes y elimina las elevaciones que nos inducirían a errores en una interpretación espectral de la voz. Esta función es la que empleamos para realizar una extracción automática de formantes, aunque para la representación espectral aplicaremos más transformaciones con el fin de mejorar la visualización final de los sonidos.

Aunque la función espectral obtenida es adecuada para realizar extracción de formantes, presenta un inconveniente en la visualización espectral. Los formantes han sido realzados considerablemente en todos los casos, con lo que se pierde la información de cuales son más y cuales son menos intensos, apareciéndonos unos decibelios similares, por ejemplo, en el primer formante de una 'o' que en el tercer formante del mismo sonido, tal y como se puede apreciar en el último caso de la figura 4.23. Esto puede dar lugar a confusiones, y no permite utilizar esta medida frecuencial en la caracterización espectral de los sonidos.

La función ponderación visualizada en los ejemplos de la figura 4.21, no presenta este problema. La función espectral final se adecua a la representación visual en frecuencias bajas y medias, mientras que la función ponderación original refleja más fielmente el comportamiento de la voz en frecuencias altas.

Con el fin de unir las mejores características de ambas funciones, se ha realizado una fusión en la que en los 900 primeros Hercios se utiliza la función espectral final, y a partir de 900 Hz se utiliza una ponderación de ambas funciones. En esta ponderación de naturaleza lineal, la función espectral final tiene más peso en las bajas frecuencias, disminuyendo progresivamente hasta el final del espectro.

En los casos 1 y 2 de la figura 4.25, aparecen las dos funciones base (espectral final y función ponderada) y el resultado al que llamaremos fusión. Como se puede apreciar, en frecuencias medias, la función fusión promedia a las anteriores, acercándose más al valor de la función ponderada a medida que el espectro aumenta de frecuencia.

La interpretación de espectros de voz es una tarea muy complicada, incluso para un experto en fonética acústica. Esto es debido, entre otros, a estos dos factores:

- La variación de las pautas de voz es muy grande. El espectro de una frase cambia según quien sea el hablante, la forma con la que ha entonado la frase, la velocidad con la que la ha dicho, la corrección con la que la ha pronunciado, etc.

- En un espectro en el que se están representando tres variables, cuesta apreciar pequeñas variaciones de las señales en el espacio en caso de realizar una representación tridimensional, o pequeñas variaciones en una escala de colores o grises en el caso de ser una representación bidimensional en la que la tercera dimensión queda representada por la coloración.

El primero de los factores expuestos es una característica inherente a la naturaleza del problema, y por lo tanto difícil de evitar, sin embargo, el segundo entra en el campo de la forma en la que se va a representar la realidad, y aquí sí se pueden tomar opciones para clarificar la visualización del espectro.

Con el fin de resaltar situaciones en las que en tres dimensiones un formante no se visualiza con claridad (por ejemplo el 2° formante del caso 3° de la figura 4.21) se ha optado por hallar los máximos y los mínimos de la función espectral a representar (la fusión) y aumentar el valor de la función en torno a los máximos, a la vez que se disminuye en torno a los mínimos.

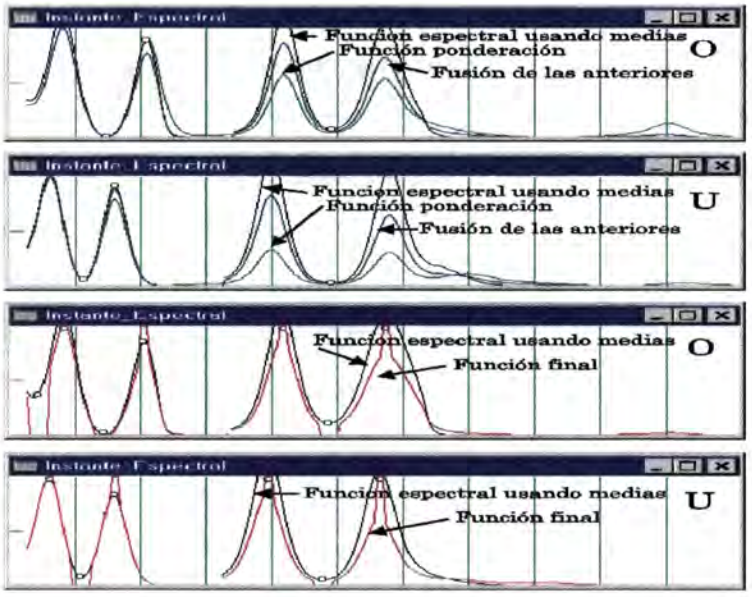


Figura 4.25

Casos 1 y 2 : Fusión de las funciones obtenidas para la obtención de un espectro representativo.

Casos 3 y 4 : Función final con los formantes realzados.

Los casos tres y cuatro de la figura 4.25 presentan sendos ejemplos comparativos de la función fusión y la función obtenida (a la que llamaremos final), elevando los máximos y disminuyendo los mínimos. Aunque en dos dimensiones esta operación parece innecesaria, en la representación tridimensional se consigue realzar y separar los formantes.

Por último, y antes de representar espectros comparativos de las funciones explicadas anteriormente, mencionar que se ha realizado un ajuste global de valores frecuenciales, de manera que se han aumentado los decibelios de las frecuencias en las que existe mayor energía de la señal de voz en el dominio del tiempo, disminuyéndose los valores en los que la energía es menor. Esta transformación se realiza debido a que todas las funciones de trabajo están normalizadas, con lo que se ha perdido la referencia de sus valores absolutos en frecuencia.

La transformación explicada en el párrafo anterior es no lineal, en ella interesa disminuir únicamente los valores de las frecuencias cuya energía se encuentre por debajo de la media. Además los valores con menor energía deben eliminarse por completo para evitar la visualización de los silencios y los sonidos sordos. La función empleada es la de la circunferencia, aplicada en el primer cuadrante, tal y como se muestra en la figura 4.24 b).

4.3.4 EVOLUCIÓN DE LOS ESPECTROS SEGÚN LAS FUNCIONES EMPLEADAS

En este apartado se compararán los espectros que se obtienen al aplicar las distintas funciones espectrales desarrolladas anteriormente.

En primer lugar se utilizarán espectros tridimensionales muy elaborados y de características idóneas, puesto que combinan una adecuada escala de colores con una visualización en el espacio dotada de vistas ocultas y curvas de nivel de los decibelios de las funciones.

De cada espectro tridimensional se mostrarán dos proyecciones, una desde un punto de vista de aproximadamente 45° en cada eje (positivo) respecto al centro de coordenadas, y otra casi de planta, que se asemeja más a un espectro bidimensional en el que la tercera dimensión se codifica mediante una escala de colores.

Para poder comparar las sucesivas mejoras que aportan las diferentes funciones espectrales, todos los espectros tridimensionales que se representan están creados a partir de una misma porción de señal de voz. Estos espectros visualizan la zona central del triptongo 'ioi'.

El primer formante de la 'i' es de menor frecuencia que el de la 'o', por lo que en el triptongo 'ioi' se aprecia en este primer formante una subida hacia la 'o' seguida de una bajada hacia la segunda 'i'. El segundo formante de la 'o' es mucho más bajo que el correspondiente de la 'i', por lo que existe una fuerte bajada hacia la 'o' seguida de una subida de igual proporción hacia la segunda 'i'. Con el tercer formante ocurre lo mismo que con el segundo, pero con una bajada más suave.

La figura 4.26 presenta la porción central del triptongo 'ioi' usando la función derivada segunda, obtenida tal y como se explicó en la sección anterior. Este es el único caso en el que la representación se realiza usando colores fríos, ya que los valores numéricos son negativos.

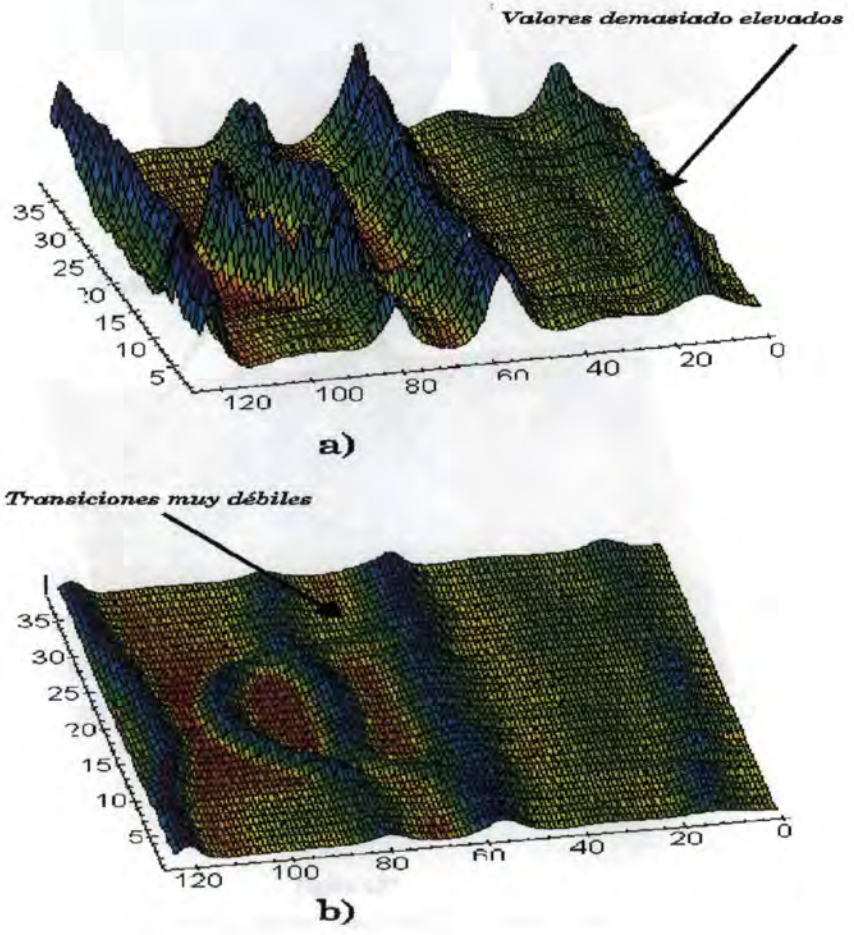


Figura 4.26

Espectro del triptongo 'ioi' usando como función la derivada segunda. Los casos a) y b) representan distintos puntos de vista de la misma figura.

El eje x de estos espectros representa las frecuencias (entre 0 y 5500 Hz), el eje z el tiempo, y el eje y los valores espectrales de señal equivalentes a los decibelios de un espectro tradicional.

En este caso el resultado presenta una buena calidad, y no existen máximos equivocados. Puesto que en el apartado anterior nos centramos en la problemática de la eliminación de los formantes erróneos, en estos espectros se pretende resaltar la capacidad que las funciones espectrales ideadas tienen para mejorar la visualización de los espectros, incluso sin acudir a sus ventajas de filtrado de máximos erróneos.

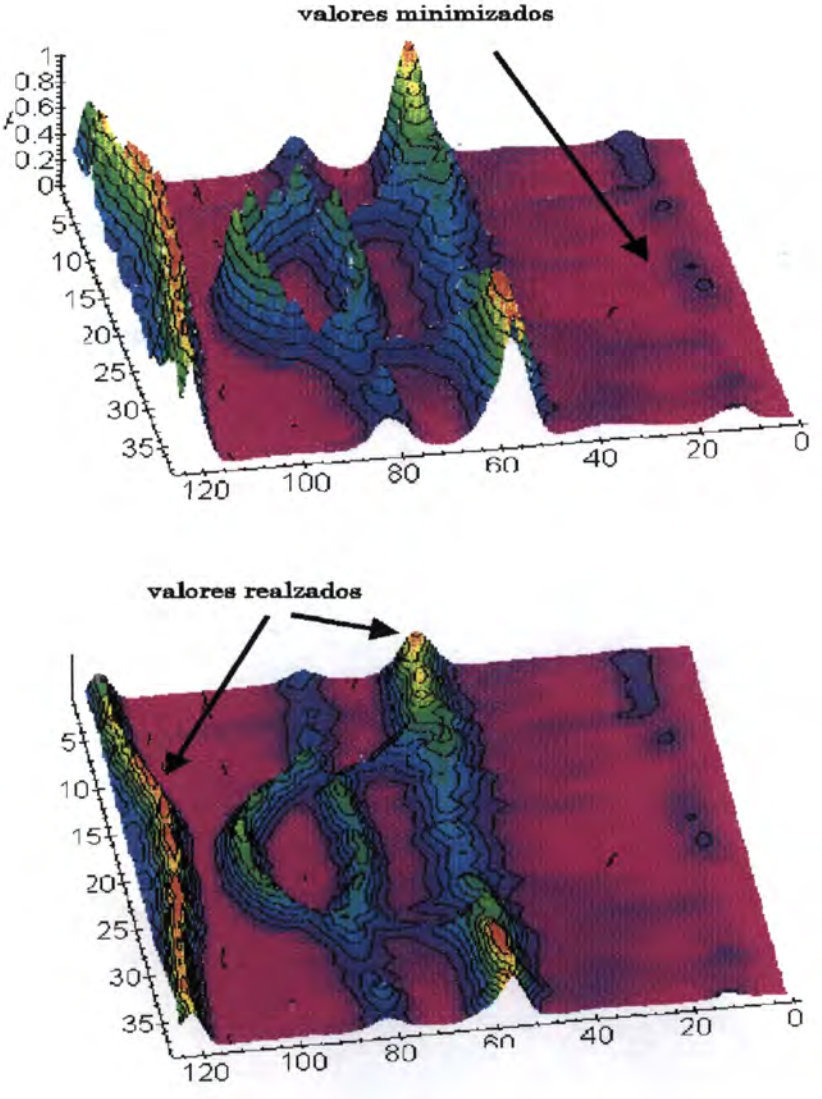


Figura 4.27
Espectro del triptongo 'ioi' usando la función ponderación.

En los espectros de la figura 4.26, cabe resaltar los valores demasiado elevados que tienen los máximos situados en altas frecuencias y que no proporcionan información útil en la caracterización de estos sonidos vocálicos. Así mismo se aprecia un debilitamiento excesivo en las transiciones del tercer formante.

En la figura 4.27 se representa el espectro obtenido a partir de la función ponderación, las frecuencias más altas del espectro han quedado atenuadas y los valores de los formantes resaltados, también las transiciones se aprecian algo mejor que en sus equivalentes obtenidas mediante la función derivada segunda.

El espectro visualizado en la figura 4.28 ha sido obtenido a partir de la función espectral que se usa para la determinación de los formantes de voz. En él se aprecia como las zonas valle son minimizadas, con el fin de evitar que pequeños máximos locales sean tomados como formantes. También los máximos se elevan algo más que con la función anterior.

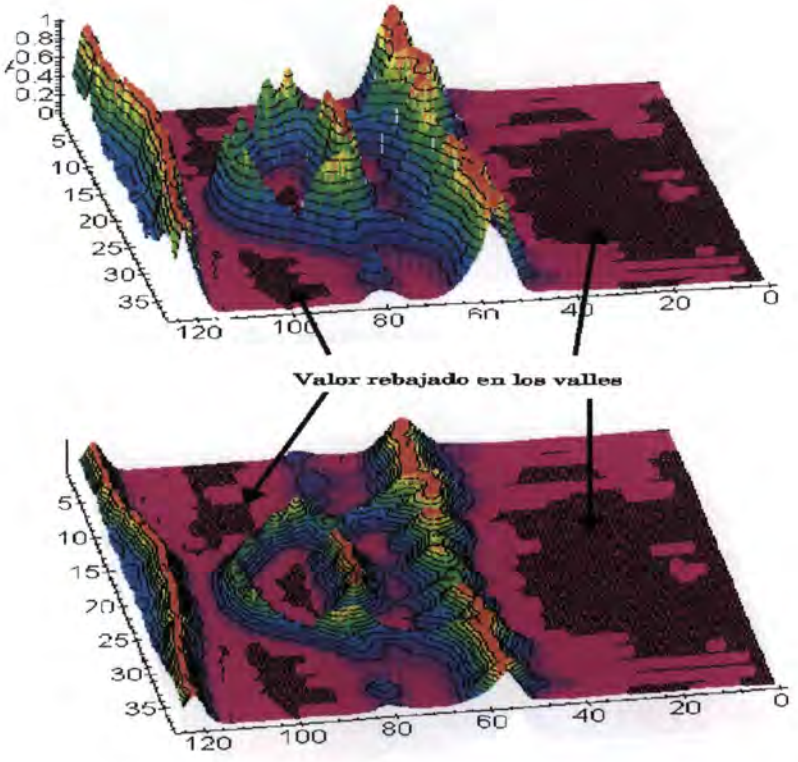


Figura 4.28

Espectro del triptongo 'ioi' obtenido usando la función espectral ponderada a partir de las medias de la función de error.

El espectro de la función de fusión en este caso no presenta ventajas apreciables, puesto que no es necesario minimizar la importancia de las frecuencias altas.

La figura 4.29 contiene el espectro calculado con la función final. Como se explicó en el apartado anterior, esta función eleva la señal en las zonas de formantes, y la rebaja en el fondo de los valles situados entre máximos relativos.

En el caso de espectros tridimensionales estas operaciones no son tan importantes, puesto que la separación entre formantes se visualiza con claridad, sin embargo, cuando se hace uso de espectros en los que la tercera dimensión se codifica con colores, las transformaciones de 'realzado' son muy útiles para diferenciar e identificar formantes.

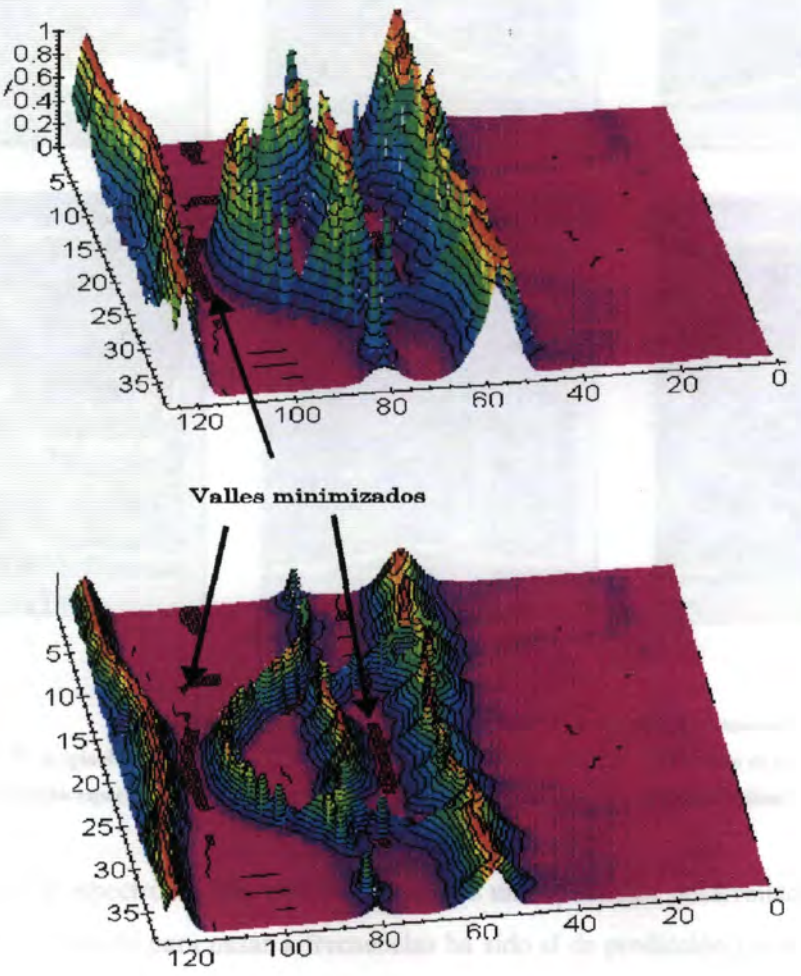


Figura 4.29

Espectro del triptongo 'ioi' obtenido usando la función final.

Los siguientes espectros que se presentan para probar las funciones empleadas son bidimensionales, codificándose la tercera dimensión mediante una escala de colores. Los colores rojo y amarillo son (en este orden) los que más decibelios representan y los que por lo tanto nos marcan la zona en la que se encuentran los formantes.

En el primer ejemplo (figura 4.30) se visualiza la secuencia 'ieaou', pudiéndose comparar los resultados según sea la función espectral empleada.

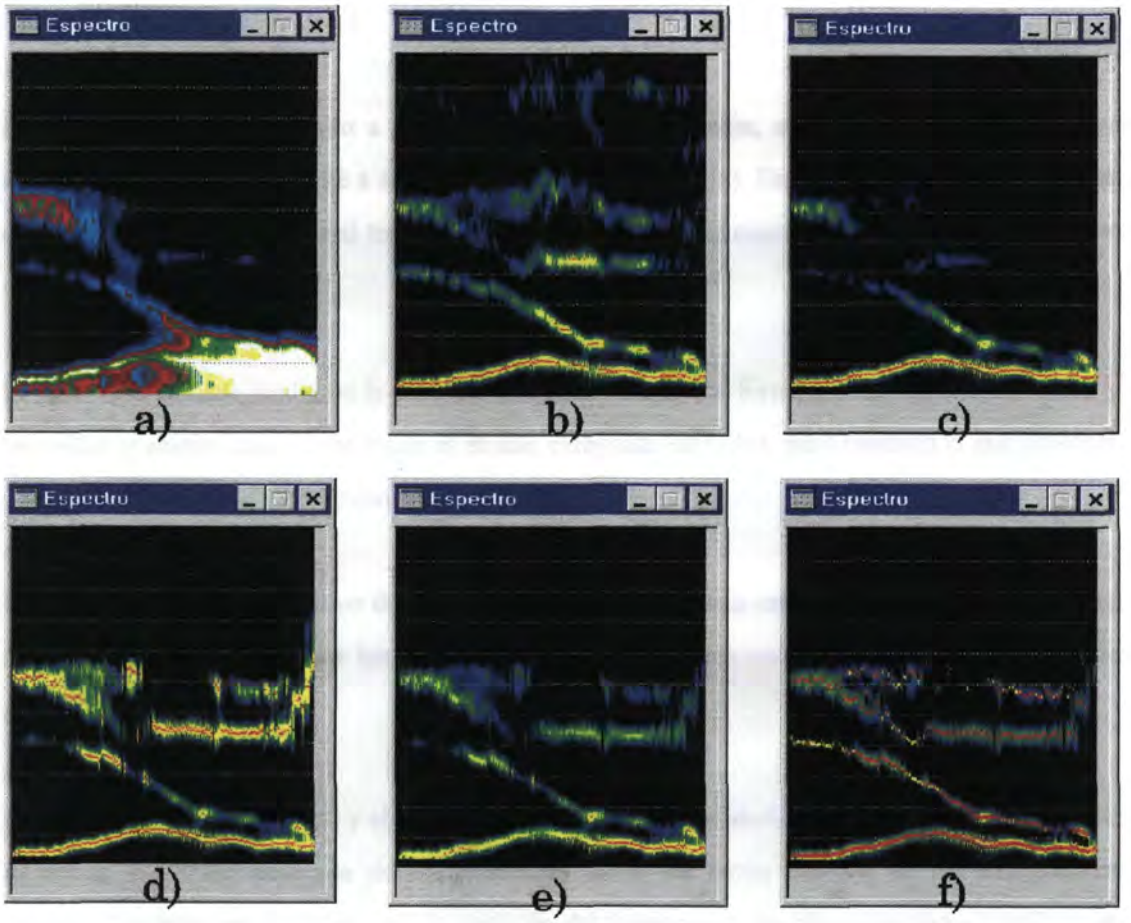


Figura 4.30

Espectros del sonido 'ieaou' usando diferentes funciones para su obtención.

a) LPC original **b) Función derivada segunda** **c) Función de ponderación**

d) Función espectral usando medias **e) Función de fusión** **f) Función final**

En el caso a), el espectro ha sido obtenido mediante una aplicación desarrollada previamente a este trabajo. El método para pasar a frecuencias ha sido el de predicción lineal. En este primer caso se puede apreciar que existe saturación y los formantes no se determinan visualmente con facilidad.

En el gráfico b) se emplea la función derivada segunda. El resultado es correcto, aunque la parte inicial (correspondiente a la vocal 'i') no aparece muy clara. Los formantes tercero y cuarto tampoco se visualizan adecuadamente. En el espectro c), correspondiente a la función de unión, los valores quedan excesivamente filtrados como para poder emplearse en una representación de calidad.

En d), como cabría esperar, los formantes quedan adecuadamente resaltados en todo el ancho de banda del espectro. Estamos empleando la función espectral tomada como base para la extracción de formantes.

El espectro e) se ha formado a partir de la función de fusión, en la que las frecuencias se minimizan proporcionalmente a su altura (cercanía a 5500 Hz). En esta ocasión el resultado es muy adecuado, debido a que el tercer formante se respeta, y el cuarto, de menor importancia, se minimiza.

El caso f) se obtiene mediante la función final, por lo que los formantes se realzan en altura y los valles se minimizan. El resultado es el más elaborado de todos, pero también el que presenta un espectro más claro y completo.

El último ejemplo comparativo de este apartado sigue la misma estructura que el anterior, pero en él se pretende caracterizar los sonidos nasales. El espectro corresponde a los sonidos 'eme ene eɳe'.

La diferencia entre el caso a) y el f) es notable en cuanto a calidad, Pudiéndose observar en éste último la evolución esperada de los formantes hacia los locus teóricos de los tres sonidos nasales. También resulta muy clara la posición del formante de nasalización que se produce en el hablante alrededor de los 2000 Hz

En los casos intermedios, como de costumbre aparece más claro el d), puesto que los máximos se resaltan especialmente con esta función espectral.

Por último, a partir de este punto, en los espectros se incluirá una gráfica de la magnitud de la señal temporal y su media. Con ello se consigue facilitar la interpretación de los espectros, especialmente en la determinación de hasta donde llega cada sonido cuando éstos son muy parecidos, como por ejemplo en algunos casos de nasales con vocales.

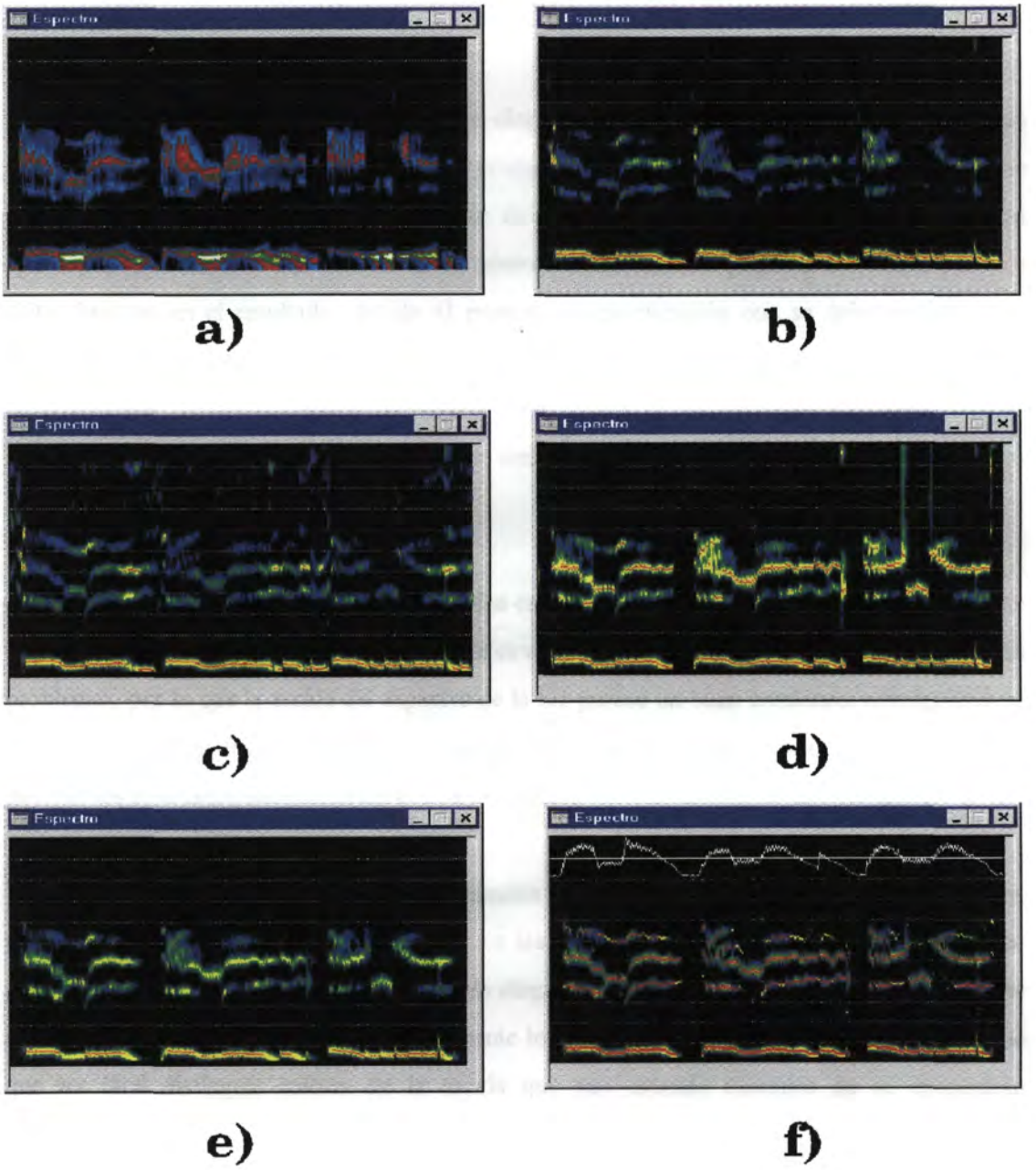


Figura 4.31

Espectros de los sonidos 'eme ene eɲe' usando diferentes funciones para su obtención.

a) LPC original **b) Función derivada segunda** **c) Función de ponderación**

d) Función espectral usando medias ***e) Función de fusión*** ***f) Función final***

4.3.5 ELECCIÓN DE LA ESCALA DE VISUALIZACIÓN DE LOS ESPECTROS

La escala de colores que se usa en los espectros es un factor de gran importancia para la correcta representación de los mismos. Una gradación de colores incorrecta produce visualizaciones inadecuadas y confusas que pueden arruinar un buen trabajo de determinación de funciones espectrales.

El primer problema que se plantea, radica en elegir el número de colores o de grises que se van a emplear. Un número demasiado elevado produce gradaciones excesivamente suaves que no permiten determinar fácilmente la evolución de los formantes, visualizándose espectros con apariencia 'difuminada'. Por otra parte, un número excesivamente pequeño de colores produce saltos bruscos en el resultado, debido al proceso de cuantización que se debe realizar para traspasar valores numéricos a colores.

El tamaño ideal de la escala se consigue con un número de entre cinco a diez colores adecuadamente elegidos.

El segundo dilema que se debe resolver se basa en la elección de un orden de colores adecuado. Copiar a la 'madre naturaleza' habitualmente es el método más elegante y eficaz de resolver los problemas, por lo que la escala del espectro de la luz parece un buen comienzo, consiguiéndose escalas de colores fríos y calientes. Desgraciadamente, en nuestro caso, estas escalas no generan los resultados satisfactorios que se esperaban.

En los espectros reales no siempre corresponden colores adyacentes a los pixels adyacentes, presentándose en muchos casos situaciones en las que una zona de un color está rodeada por pixels de dos colores de distancia en la escala elegida. Esto nos obliga a escoger escalas en las que no sólo tienen que diferenciarse claramente los colores adyacentes, sino que además tiene que ser fácil distinguir colores de la escala que aun estando cercanos no se encuentran contiguos.

En los espectros que se visualizan con la herramienta desarrollada, se han elegido varias combinaciones de colores posibles. Habitualmente usamos una escala con fondo negro, colores fríos al comienzo, calientes al final y acabando con el rojo. Los espectros visualizados en este apartado corresponden a la configuración descrita.

En la figura 4.32 se presentan los espectros de 'ieaou' y 'eme ene eɳe' utilizando otras tres escalas posibles. En el primer caso, únicamente se cambia el fondo negro por el blanco. Esto es muy útil para imprimir los espectros de una forma rápida y económica, manteniendo el esquema de colores al que estamos acostumbrados. El mayor problema es que los colores azules (menos significativos) obtienen sobre el blanco un relieve que no les corresponde.

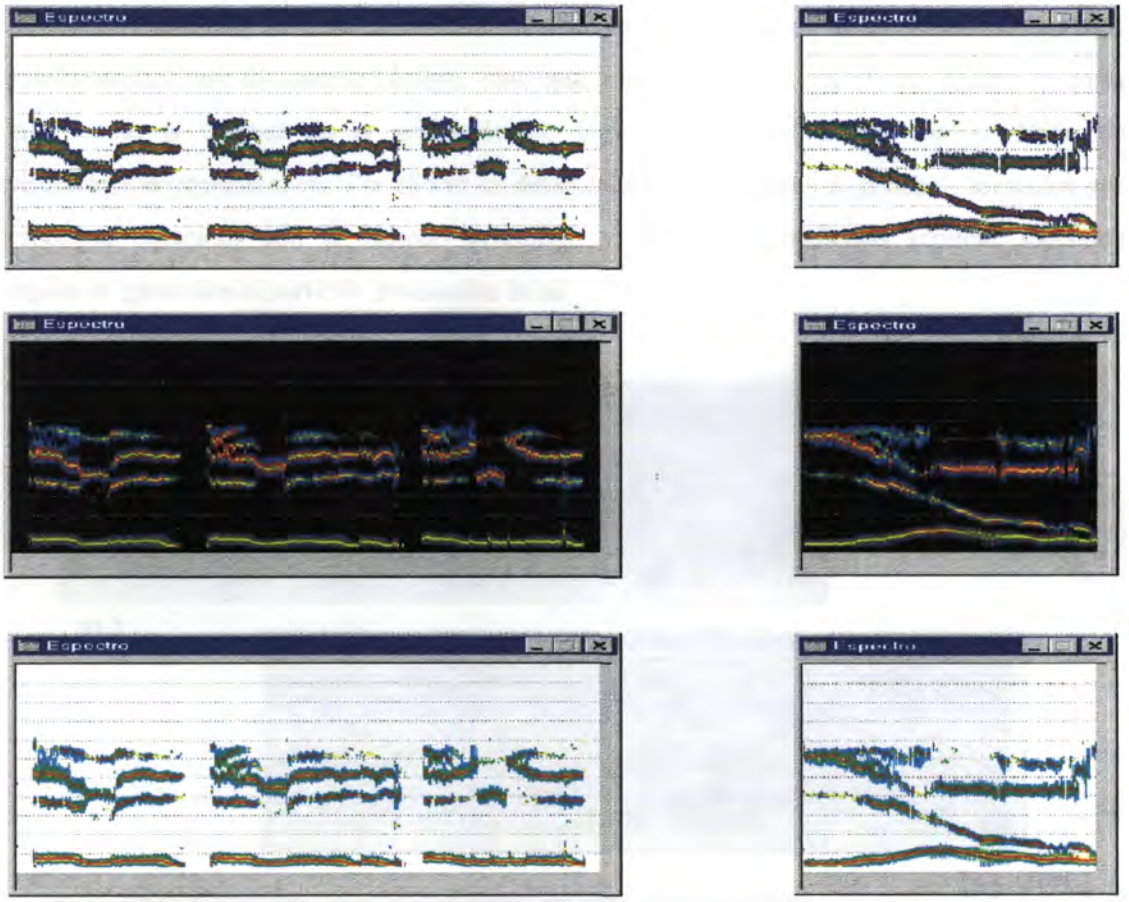


Figura 4.32
Ejemplos de espectros usando distintas escalas de colores.
izquierda: 'eme ene eɳe' derecha: 'i e a o u'

El segundo caso (fondo negro) se basa en hacer corresponder los lugares más significativos de la escala con los colores más luminosos (amarillo y blanco).

Los últimos espectros de la figura 4.32 sirven de ejemplo de como no debe diseñarse una escala. Aquí, el problema reside en que los colores amarillo, verde claro, azul claro y blanco no se distinguen con facilidad, por lo que si los decibelios obtenidos no llegaran al color rojo, sería difícil distinguir la importancia de cada parte del espectro.

4.3.6 RESULTADOS

Una vez explicados y comparados los distintos algoritmos y funciones matemáticas ideados para la obtención de los espectros, en este apartado se presentarán casos de ejemplo de espectros finales de sonidos sonoros.

La claridad de un espectro no depende únicamente de la idoneidad de los procesos matemáticos e informáticos con los que se obtiene, sino que además influyen significativamente factores tales como la calidad de la grabación de partida, la precisión del soporte físico de visualización/impresión, etc. Por ello se ha decidido establecer 'pares de prueba', formados por espectros obtenidos con el método propuesto y sus correspondientes réplicas mediante espectros generados a partir de predicción lineal.

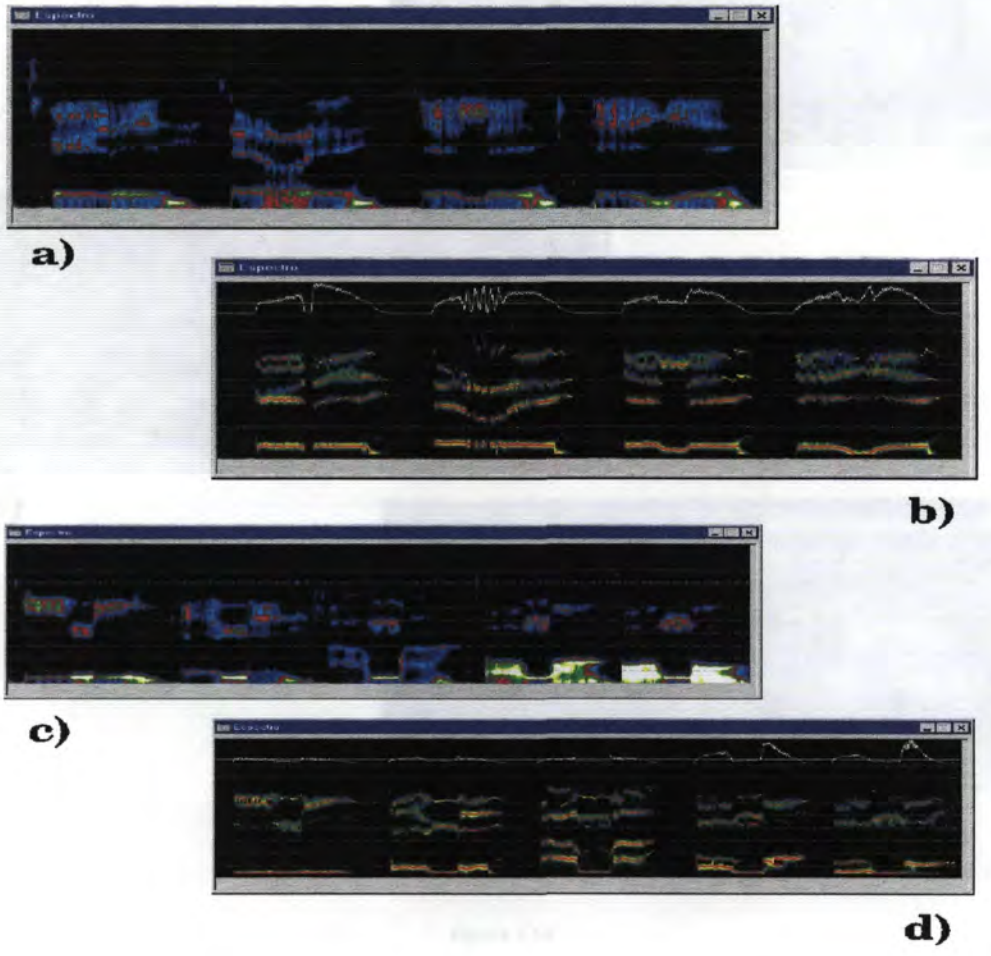


Figura 4.33
Sonidos 'ere erre ele eλe' y 'imi eme ama omo umu' empleando los espectros propuestos y sus correspondientes LPC

La figura 4.33 presenta en primer lugar los espectros tradicional LPC en a) y final en b), pertenecientes a la secuencia de consonantes sonoras vibrantes y laterales: 'ere erre ele eλe'. Aunque en el espectro LPC se aprecian las características básicas de estos sonidos, en el caso b) la claridad es bastante mayor, y los formantes aparecen más nítidos y delimitados.

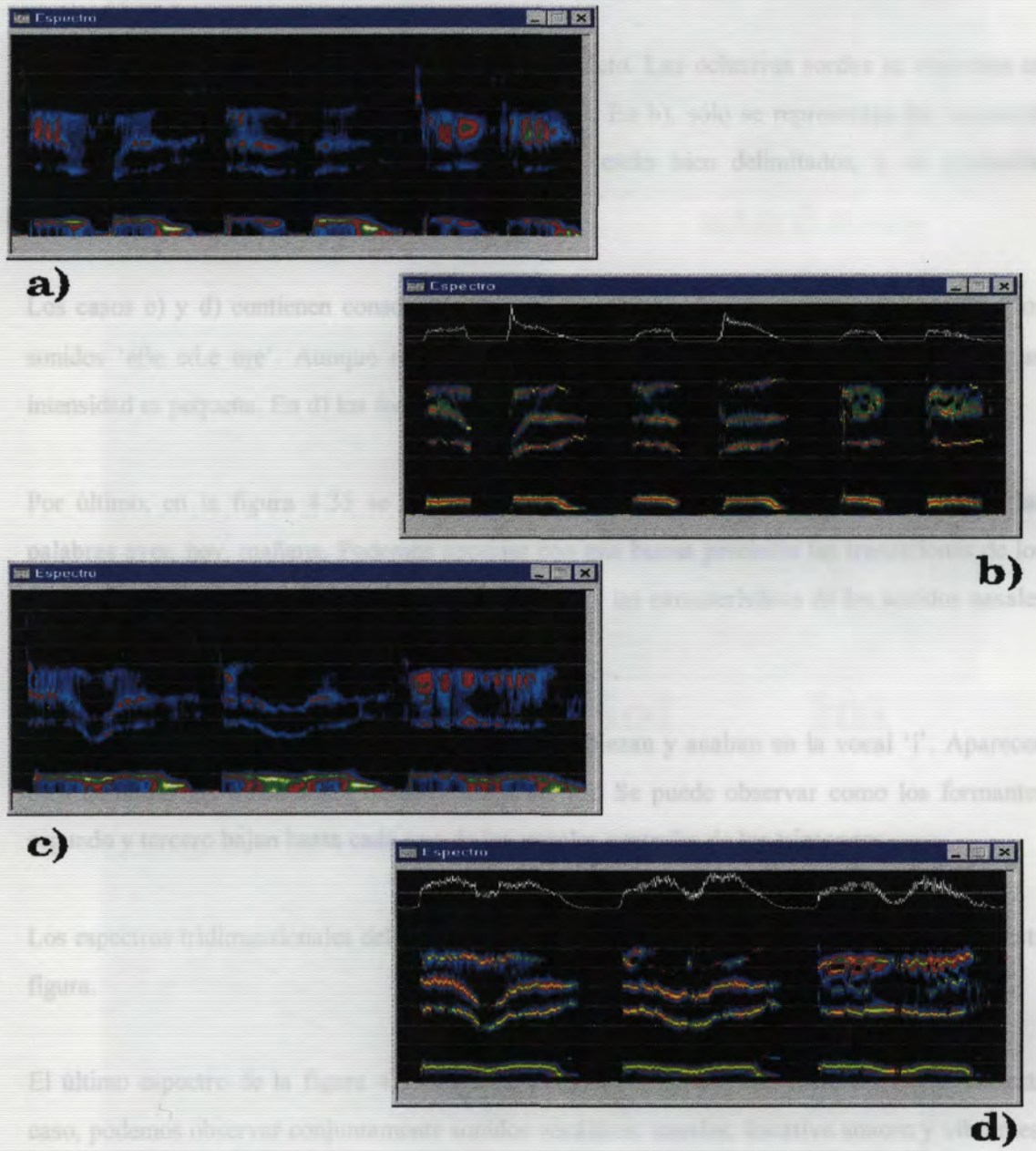


Figura 4.34

Sonidos 'epe ete eke' y 'eβe ed.e eγe' empleando los espectros propuestos y sus correspondientes LPC

Los casos c) y d) se corresponden con la secuencia de sonidos 'imi eme ama omo umu'. En el caso d) se evita la saturación de la escala y se delimitan más claramente los formantes de las vocales. Las posiciones frecuenciales de la nasal quedan bien definidos en ambos casos.

La figura 4.34 se centra en el grupo oclusivo completo. Las oclusivas sordas se muestran en los casos a) y b) mediante los sonidos 'epe ete eke'. En b), sólo se representan las secciones sonoras. Como se puede apreciar, los formantes están bien delimitados, y su evolución corresponde con la esperada.

Los casos c) y d) contienen consonantes fricativas sonoras. La grabación corresponde a los sonidos 'eβe ed.e eγe'. Aunque en el caso c) la trayectoria de los formantes es clara, su intensidad es pequeña. En d) los formantes aparecen suficientemente realzados.

Por último, en la figura 4.35 se presentan tres espectros, el primero de ellos contiene las palabras ayer, hoy, mañana. Podemos apreciar con una buena precisión las transiciones de los formantes al evolucionar de una vocal a otra ('hoy') y las características de los sonidos nasales en 'mañana'.

El segundo espectro contiene los triptongos que empiezan y acaban en la vocal 'i'. Aparecen bien definidas las transiciones desde y hasta las ies. Se puede observar como los formantes segundo y tercero bajan hasta cada una de las vocales centrales de los triptongos.

Los espectros tridimensionales del apartado anterior se corresponden con el sonido 'ioi' de esta figura.

El último espectro de la figura 4.35 muestra como se visualiza una frase completa. En este caso, podemos observar conjuntamente sonidos vocálicos, nasales, fricativo sonoro y vibrantes. La evolución de los formantes aparece muy nítida en todo el intervalo mostrado. Las nasales se distinguen bien y la vibrante múltiple se aprecia con claridad.

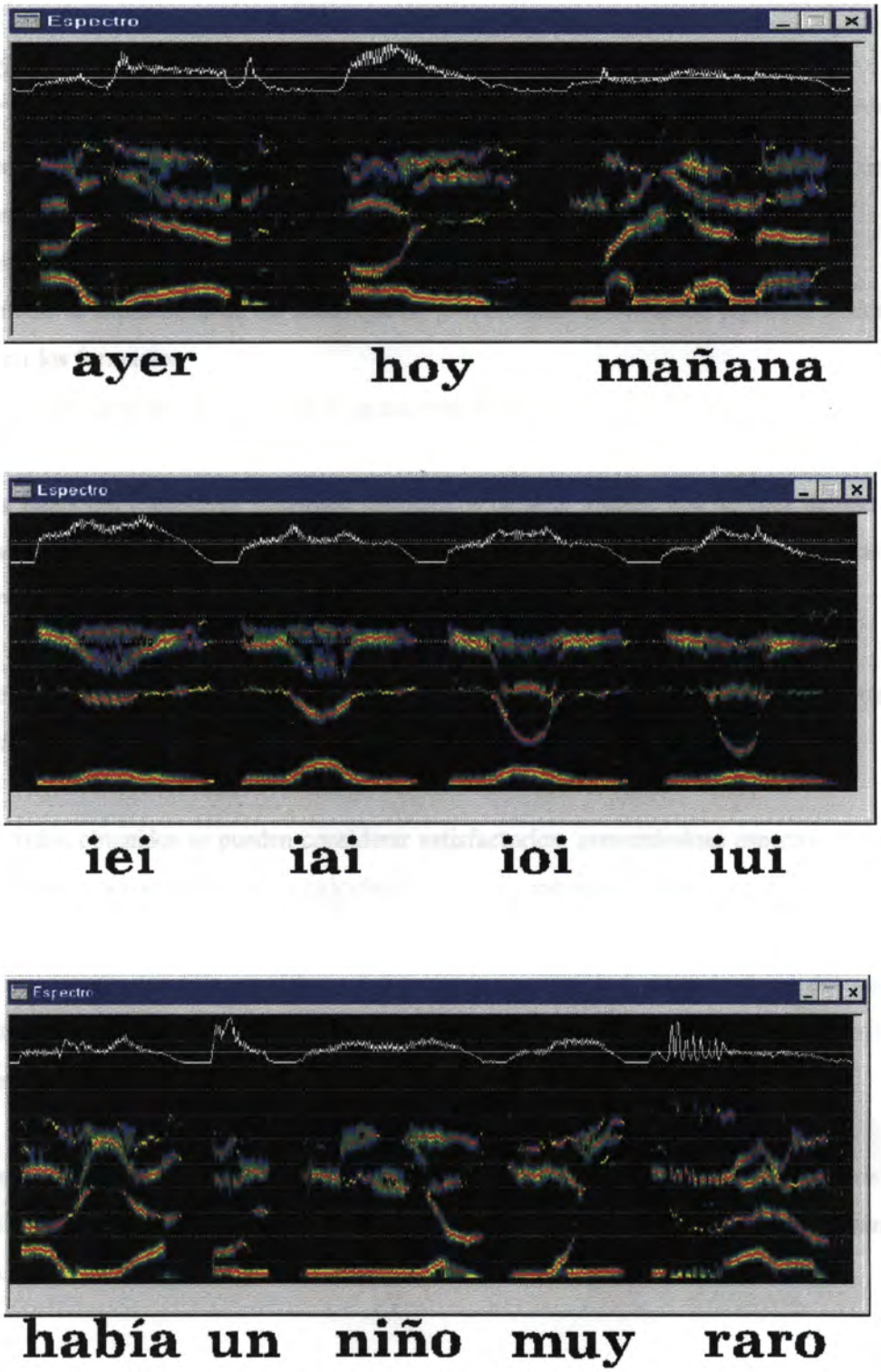


Figura 4.35

Espectros finales utilizando el método propuesto

4.3.7 CONCLUSIONES

La obtención de espectros claros, requiere la utilización de varios filtros que realicen las siguientes funciones:

1. Eliminar formantes erróneos.
2. Disminuir la importancia de las zonas que no proporcionan información espectral básica.
3. Realzar los formantes.
4. Minimizar los valores de las zonas situadas entre formantes.

Entre las funciones espectrales obtenidas existen dos de gran interés, una de ellas debido a su idoneidad en la extracción automática de formantes, y la otra por sus especiales características para la representación visual de los espectros de voz.

La elección de una escala de colores adecuada es muy importante. Esto es debido a que nos permitirá identificar con mayor facilidad las zonas más representativas del espectro.

Los resultados obtenidos se pueden considerar satisfactorios, presentándose espectros bastante más claros que sus correspondientes calculados con otros métodos y herramientas.

Como único punto negativo, se debe señalar que la complejidad computacional del método propuesto es bastante superior a la que conllevan los algoritmos tradicionales de creación de espectros.

Para completar este trabajo, se debe investigar en mecanismos de obtención de características espectrales en sonidos sordos, y este será el objetivo a cubrir en la siguiente sección del capítulo.

4.4 DETERMINACIÓN DE CARACTERÍSTICAS ESPECTRALES EN SONIDOS SORDOS

4.4.1 RESUMEN

La naturaleza de los sonidos sonoros varía considerablemente de la de los sordos, por lo que los métodos y algoritmos empleados en unos no son completamente utilizables para los otros, por ello ha sido necesario adaptar estos algoritmos para que reflejen adecuadamente las nuevas características espectrales mostradas por los sonidos sordos.

La necesidad de utilizar distintos métodos matemáticos y algorítmicos en los sonidos sordos y sonoros, ha obligado a idear una manera automática de distinguir ambos tipos de sonidos.

Al igual que en las secciones anteriores de este capítulo, se presentan diversas gráficas y espectros comparativos que ayudan a la comprensión de las decisiones tomadas y los algoritmos ideados.

4.4.2 INTRODUCCIÓN

Los sonidos sordos de la lengua castellana corresponden a la mayor parte de las fricativas, el sonido africado y las oclusivas sordas, en total: [p], [t], [k], [tʃ], [f], [θ], [s] y [x].

La determinación espectral de las oclusivas sordas se realiza en gran medida estudiando la evolución de los formantes de los sonidos sonoros que las rodean [MOR90], [MUJ90], sin embargo, la obtención de las características de la barra de oclusión (anchura y posición frecuencial con los mayores decibelios) requieren de un estudio y algoritmos específicos [BON96], [BLU80].

Los sonidos fricativos y el africado contienen una estructura mucho menos estable y periódica que los sonidos sonoros [SMI94], por lo que habrá que centrarse en la localización frecuencial

del ruido que contienen. En este caso, la búsqueda de formantes claros y representativos no tiene el sentido que existe en la determinación de los sonidos sonoros.

Aunque la claridad de un espectro depende fundamentalmente de la precisión obtenida en los segmentos sonoros, una buena representación de los sonidos sordos puede ayudar considerablemente a conseguir una adecuada interpretación de las señales visualizadas mediante espectros de voz.

En los siguientes apartados se explicará la manera con la que se han distinguido de forma automática los sonidos sordos de los sonoros, las funciones que se consideran adecuadas para caracterizar los sonidos sordos, y la forma de representar los resultados. También se muestran espectros obtenidos a lo largo del estudio y espectros finales fruto de los resultados combinados en la representación de los sonidos sordos y sonoros.

4.4.3 DETECCIÓN DE LA SONORIDAD-SORDEZ

Los espectros de voz tradicionales se obtienen aplicando operaciones matemáticas a las señales sonoras temporales, de tal forma que se calculen sus transformadas frecuenciales y posteriormente se visualice el resultado. Este modo de operar no requiere de ningún tipo de reconocimiento de la señal que se está analizando.

En nuestro caso, el modo de actuar se complica, puesto que a los resultados frecuenciales se les aplica una nueva etapa de "realzado" de las características espectrales de la señal de voz. Puesto que vamos a tener dos algoritmos diferentes de realzado (uno para los sonidos sonoros y otro para los sordos), resulta necesario disponer de algún método de decisión de cuando aplicar uno u otro, por lo tanto, deberíamos ser capaces de reconocer automáticamente la característica de sonoridad-sordez. Los parámetros básicos en los que nos basaremos son los siguientes:

1.- Energía de la señal
   Los sonidos sonoros generalmente presentan más energía que los sordos.
2.- Cruces por cero. Máximos
   Los sonidos sonoros, generalmente tienen menos cruces por cero y máximos que los sonidos sordos.
3.- Barra de sonoridad
   Los sonidos sonoros, generalmente presentan un armónico de gran intensidad en una frecuencia muy baja.

Las dos primeras características analizan la señal en el dominio del tiempo, mientras que la tercera lo hace en el de la frecuencia. Habitualmente se ha tomado la barra de sonoridad como el parámetro más fiable de detección de la sonoridad-sordez, sin embargo, ninguno de los tres parámetros resulta totalmente seguro.

Después de realizar diversas pruebas de cada uno de los parámetros aisladamente y de combinaciones entre ellos, se ha llegado a la conclusión, en este trabajo, de que lo más fiable es utilizar una combinación de los tres métodos, con el fin de dotar de robustez al algoritmo final de detección.

La figura 4.36 presenta un ejemplo en el que se combinan vocales, silencios, todos los sonidos fricativos y el africado.

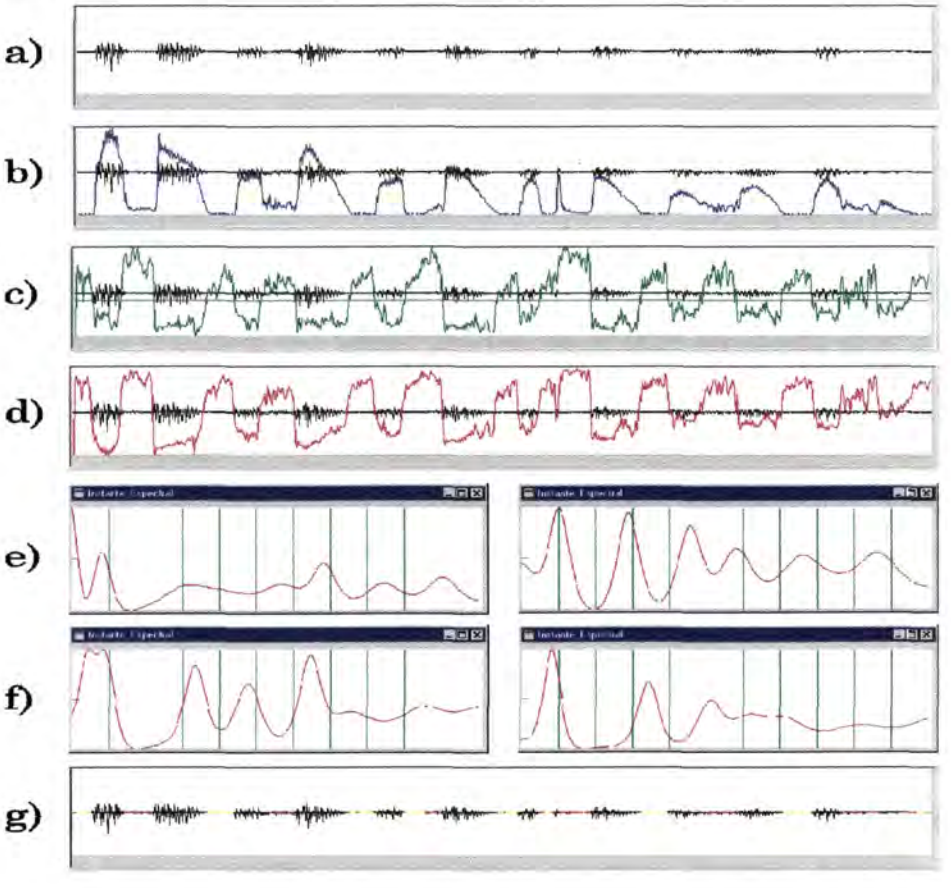


Figura 4.36

Características de la señal de voz para la detección de la sordez-sonoridad.

El gráfico a) presenta la señal en voz en tiempo de la que se parte. En b) se representa la energía de la señal ponderada con su media aritmética. En este caso, la energía de la señal indica bastante bien la localización de los sonidos sordos y sonoros, aunque no de forma

perfecta, por ejemplo, falla en la última vocal del fichero, debido a la pérdida de energía que se produce al acabar la frase.

La energía de la señal presenta el problema de que si se establece un umbral de decisión (por debajo del umbral => sonido sordo, por encima => sonoro), la decisión fallará a menudo, debido a que el volumen con que se realizan las grabaciones no se puede mantener constante (varía según el hablante, el micrófono, la configuración de la tarjeta digitalizadora, etc.). La alternativa consiste en normalizar el volumen de la señal de voz de entrada, y esta es la solución que aquí se ha adoptado.

En c) se representan los máximos de la señal de voz. Aunque hallar los máximos computacionalmente es más complejo que los cruces por cero, presenta la ventaja de eliminar el problema de la componente continua en la señal, que a veces falsea los resultados. Como se puede observar, los máximos se reducen en las vocales y aumentan considerablemente en los silencios y sonidos sordos.

En d) se ha dibujado la función matemática que pondera a partes iguales la energía (de forma inversa) y los máximos, con lo que se obtiene una función que indica la sordez de la señal. A la vista del gráfico se observa que cuando esta función se sitúa por encima de su media, casi siempre se corresponde con una porción de sonido sordo.

Los casos e) y f) muestran instantes espectrales de la señal correspondientes a sonidos sordos y sonoros respectivamente. Las gráficas corresponden a la derivada segunda de la señal evaluada en $r=0.9$, que se utiliza en la mayor parte de los algoritmos presentados en este capítulo.

Los sonidos sordos casi siempre presentan una pendiente negativa en las primeras frecuencias del espectro, a diferencia de los sonidos sonoros, que comienzan habitualmente con una subida que les lleva hacia el máximo correspondiente a su primer formante. Además, la energía correspondiente a las primeras frecuencias analizadas suele ser mayor en los sonidos sordos.

Uniendo todas las características mostradas y estableciendo empíricamente valores umbral de decisión para cada uno de los métodos, se ha creado un algoritmo de detección de sordez-sonoridad que resuelve satisfactoriamente la gran mayoría de los casos que se presentan.

El gráfico g) corresponde a la señal temporal de voz con la característica de sordez resaltada sobre el eje horizontal. Este es un resultado típico que muestra como sin obtenerse resultados perfectos, el método sirve adecuadamente al objetivo final de representación de espectros de voz.

4.4.4 FUNCIONES ESPECTRALES DESARROLLADAS

El primer paso que deberemos dar es la evaluación de resultados de los espectros LPC tradicionales, con el fin de determinar si la calidad de los mismos es suficiente, y si no lo es, en qué casos se debería mejorar.

La figura 4.37 muestra el espectro LPC de la secuencia "ese efe etʃe exe eθe", que será utilizada a lo largo de esta sección para ilustrar buena parte de los conceptos que aquí se presentan.

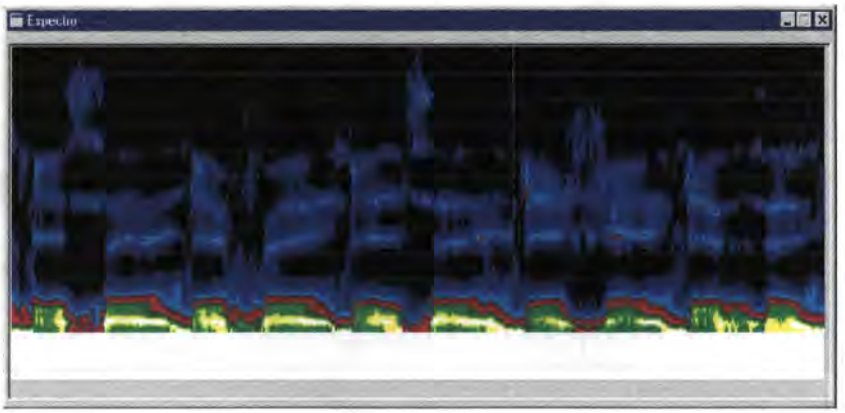


Figura 4.37
Espectro LPC tradicional de la secuencia "ese efe etʃe exe eθe".

A la vista del espectro, se puede apreciar que las características de los sonidos [s] y [tʃ] se muestran con suficiente claridad debido a sus porciones de ruido en frecuencias altas, sin embargo, [x], y especialmente [f] y [θ] no presentan características claras que las determinen, esto es debido a que la mayor parte de su energía se encuentra concentrada en frecuencias bajas. Las frecuencias medias y altas, donde se encuentra información de importancia no se representan con claridad debido a su baja intensidad.

En segundo lugar analizaremos las funciones espectrales básicas de las que partimos para la caracterización de los sonidos. La figura 4.38 presenta cuatro casos de instantes espectrales de sonidos sordos. Las curvas se corresponden con los polos evaluados en r=0.9 (color verde) y su derivada segunda (color rojo).

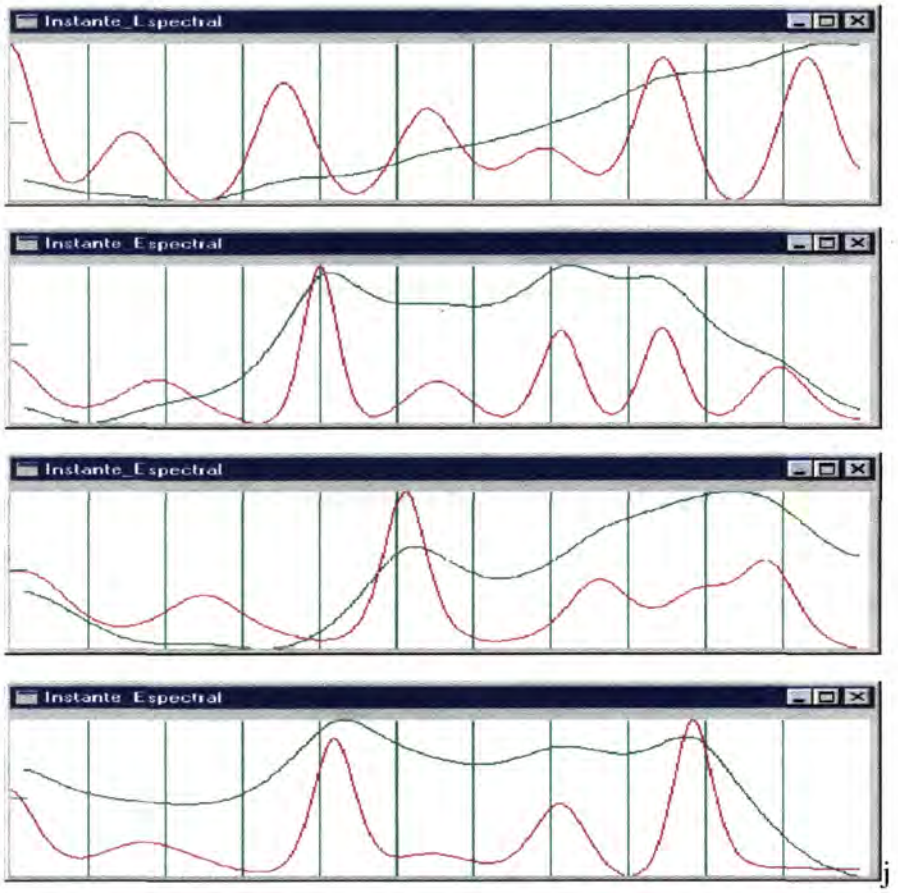


Figura 4.38
Polos y derivada segunda de cuatro instantes espectrales de diferentes sonidos sordos.

Resulta interesante observar como la función que mejor determina el tipo de sonido que representa, es en estos casos la correspondiente a los polos. Esto no es sorprendente, puesto que a diferencia de los sonidos sonoros, aquí la información de importancia se consigue con la determinación de las zonas frecuenciales con más energía, no con la evolución concreta de la variación de la señal en estas zonas, que en el caso de los sonidos sonoros nos daba lugar a los formantes.

Tomando como ejemplo el primer caso (gráfico superior) de la figura 4.38, correspondiente a un instante del sonido fricativo [s], observamos que los decibelios aumentan con la frecuencia,

y que esta proporción se mantiene hasta el final del espectro. Esto corresponde adecuadamente con las características esperadas del sonido analizado.

La información de "detalle" proporcionada por la derivada segunda, no nos ayuda a clarificar la naturaleza del sonido, puesto que se rige por las fluctuaciones (no significativas en sus detalles) del ruido de [s].

En definitiva, la función de minimización de error (polos) resulta muy adecuada para caracterizar los sonidos sordos. Por desgracia, una cosa es caracterizar instantes espectrales y otra muy diferente visualizar las señales de voz. El problema surge de la combinación de estos hechos:

1. La función de polos es muy diferente a su derivada segunda.
2. El algoritmo de determinación de la sordez-sonoridad de los sonidos, en ocasiones deja "islas" de sonidos sordos en medio de los sonoros y viceversa.

Debido a los factores anteriores, si empleáramos la función de polos para visualizar los sonidos sordos, nos encontraríamos con espectros en los que existirían fuertes discontinuidades en instantes de tiempo aislados. La función derivada segunda no presenta este problema, pero carece de las cualidades de la función de polos.

La solución adoptada se basa en crear una nueva función que pondere las anteriores. La figura 4.39 presenta las gráficas correspondientes a los polos (verde), función de ponderación (azul) y otra (negro), en la que se elevan los polos al cuadrado con el fin de acercar la solución a estos valores, pero sin alejarlos excesivamente de la derivada segunda, para evitar discontinuidades en la visualización de los espectros.

En general, se consigue el efecto de que si los polos fueran la envolvente de la función derivada segunda, la función final conseguiría aproximar mejor la envolvente original. Este efecto se aprecia muy bien en el primer caso de la figura 4.39.

A partir de ahora, utilizaremos en las gráficas comparativas la función de polos y la función final a la que llamaremos función de aproximación.

Otro problema importante reside en la necesidad de visualizar los sonidos, pero dejar vacíos los espacios espectrales correspondientes a los silencios. Para hacer esto de una manera perfecta,

habría que ser capaces de distinguir unos de otros, y esto no resulta sencillo. Habitualmente esta cuestión se soluciona en las representaciones espectrales, poniendo un umbral de visualización, de tal forma que se asume que los valores por debajo del umbral corresponden al silencio o a valores frecuenciales poco importantes de los sonidos. Normalmente esto funciona bien, pero lógicamente siempre se darán situaciones en las que se visualicen los silencios o en las que no se visualicen realizaciones de baja intensidad, normalmente correspondientes a los sonidos sordos.

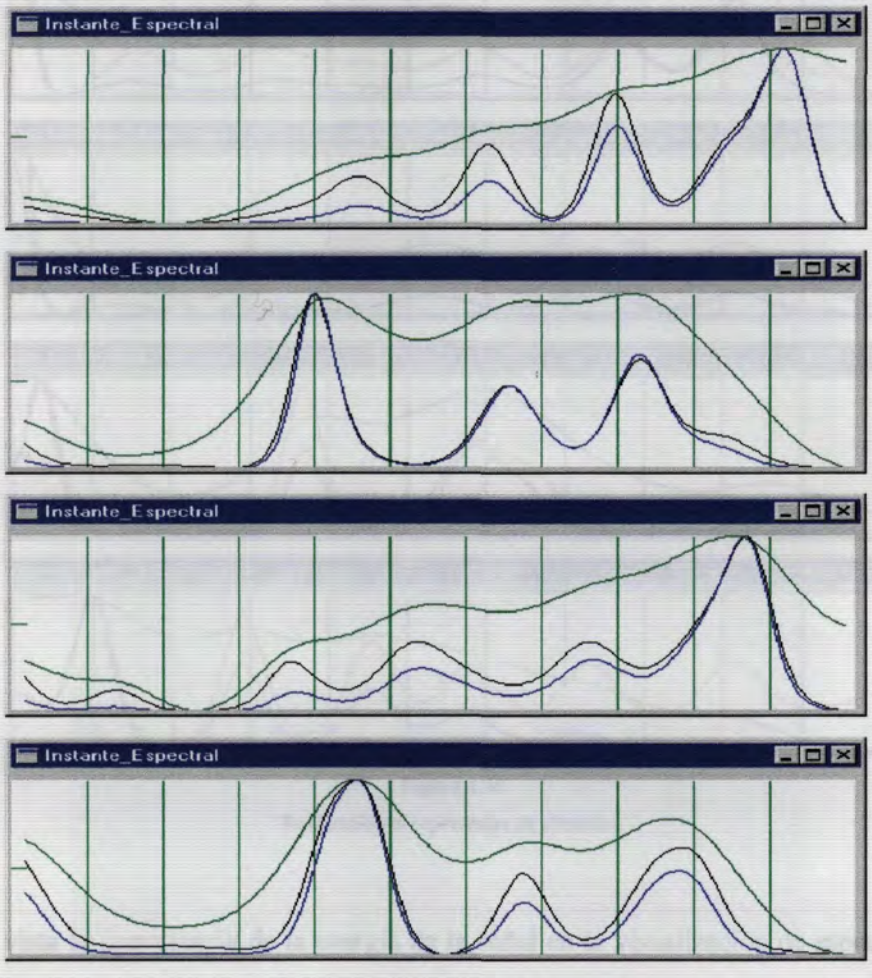


Figura 4.39

Polos, función ponderación y función de aproximación de cuatro instantes espectrales de varios sonidos sordos.

La figura 4.40 presenta cuatro instantes espectrales correspondientes a silencios. Los dos primeros han sido tomados de los espacios "vacíos" de sonidos oclusivos, y los dos últimos de silencios existentes entre palabras de una frase. Las funciones son las explicadas en la figura anterior.

Como se puede observar, en las dos últimas gráficas la evolución frecuencial es muy parecida a la de los sonidos sonoros, esto es debido a la influencia que ejercen las señales sonoras que rodean al silencio. Esta característica hace muy difícil distinguir los silencios en el dominio de la frecuencia sin recurrir a la intensidad de la señal.

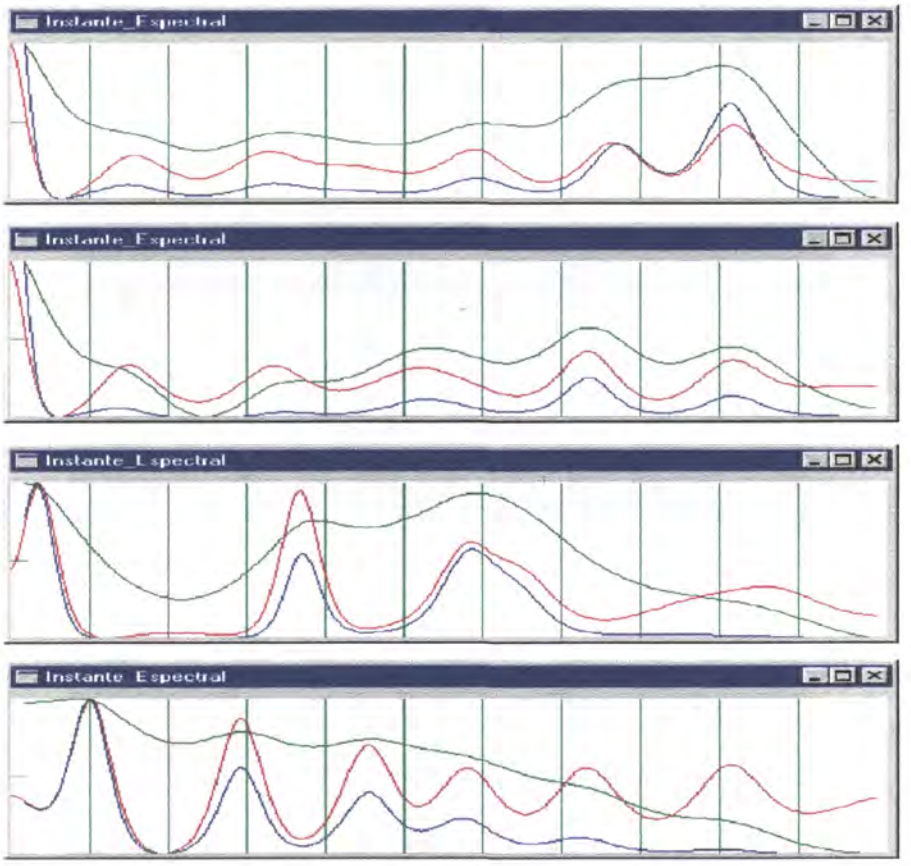


Figura 4.40

Realizaciones espectrales de silencios.

Una vez vista la importancia de la energía de la señal en la visualización de espectros de voz, estableceremos un factor de intensidad que nos corrija los valores de amplitud de los instantes espectrales, con el fin de filtrar las zonas de silencios.

En la figura 4.41 se presentan a) las zonas de correspondientes a sonidos sordos de la señal analizada, b) el espectro usando la función de aproximación sin establecer factor de intensidad, c) el caso anterior estableciendo factor de intensidad. Como se puede apreciar, la introducción de este parámetro consigue eliminar las zonas de silencios, respetando los sonidos más sensibles a este tipo de filtrados. A partir de ahora, todos los espectros se visualizarán utilizando el factor de intensidad explicado.

La figura 4.42 contiene los espectros de voz obtenidos aplicando la función de polos (gráfico superior) y la función de aproximación (gráfico inferior).

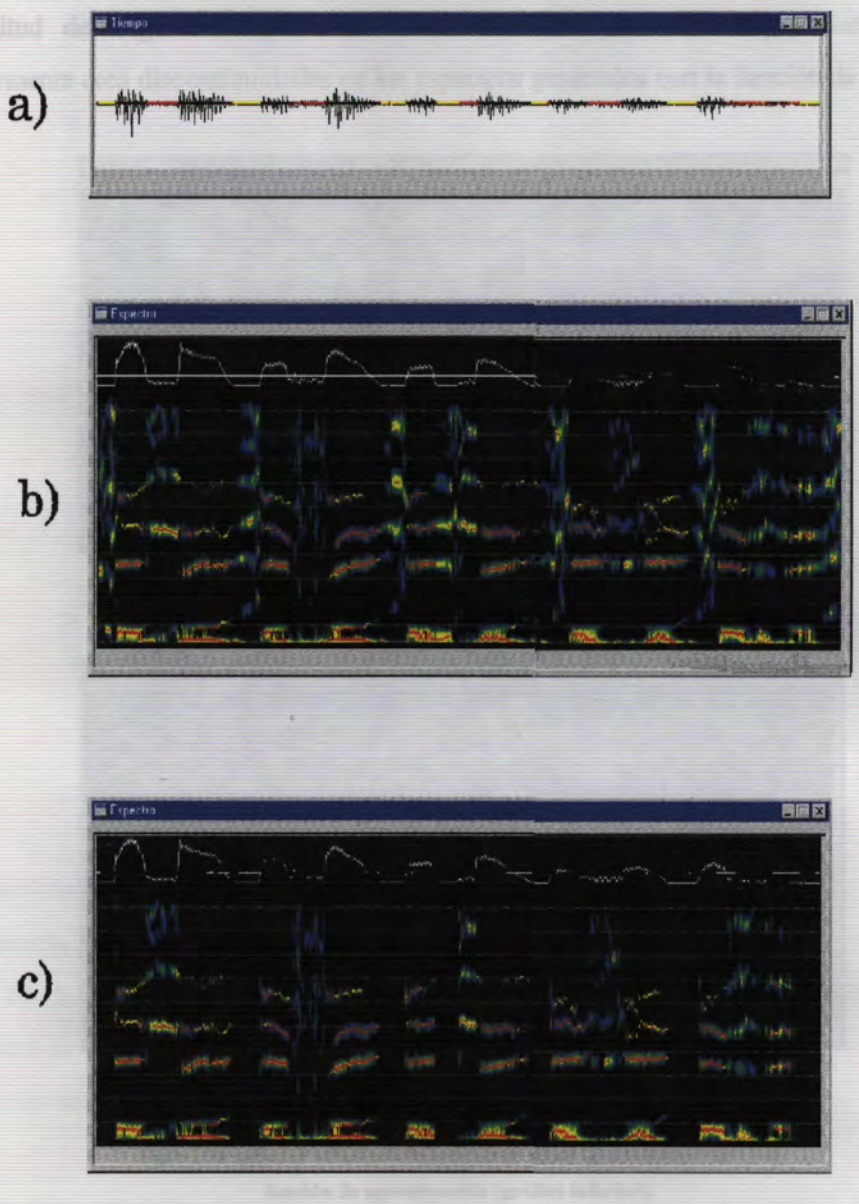


Figura 4.41

Resultado de la aplicación del factor de intensidad sobre un espectro de voz.

En este caso, los umbrales de visualización han sido colocados deliberadamente bajos, con el fin de resaltar el siguiente concepto: puesto que las funciones espectrales se normalizan, en el caso de existir grandes diferencias en los decibelios hallados en las distintas frecuencias, solamente se visualizan las de mas intensidad. Esta situación no es conveniente y se produce

habitualmente en los sonidos sordos. Resulta adecuado aplicar un filtro paso bajo que elimine las bajas frecuencias con altas energías.

La gráfica superior de la figura 4.42 muestra las "rayas" que se producen debido a la inexactitud del algoritmo de determinación de sordez-sonoridad, que como se explicó anteriormente crea discontinuidades en los espectros generados con la función de polos.

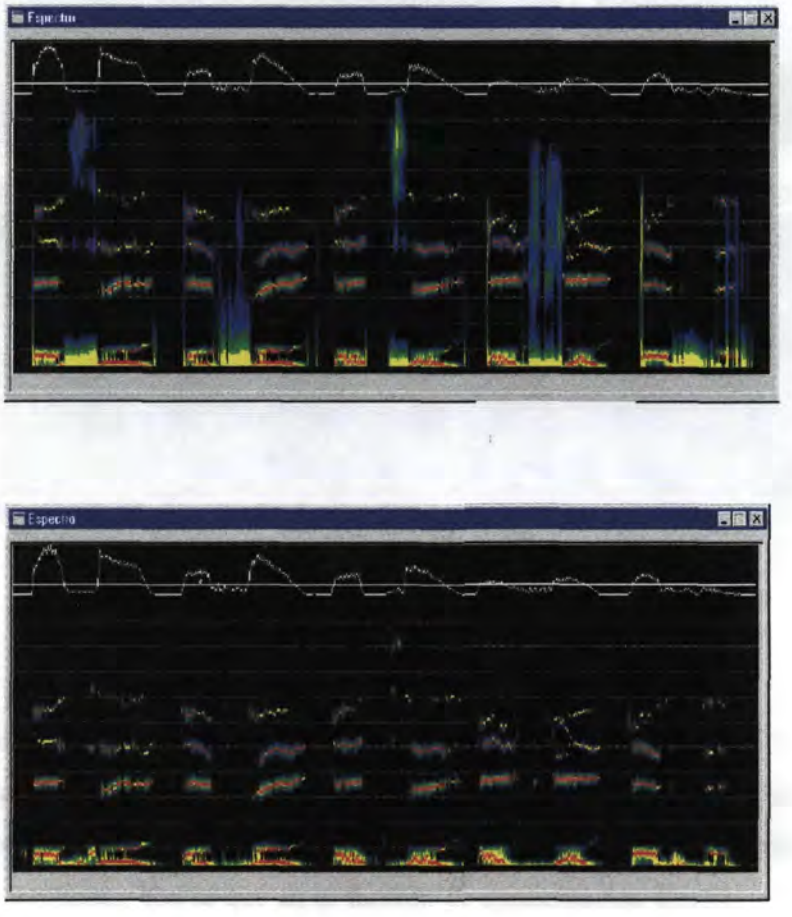


Figura 4.42
Espectros hallados utilizando la función de polos (gráfico superior) y la función de aproximación (gráfico inferior).

La figura 4.43 se corresponde con la anterior, pero aplicando un filtro paso bajo. En este caso, los "vacíos" en bajas frecuencias de los sonidos sordos han sido forzados, y a cambio, las características espectrales de las demás frecuencias se intensifican en los gráficos.

La determinación de las características espectrales de los sonidos sordos, resulta más sencilla en el espectro generado a partir de los polos, tal y como se había previsto inicialmente.

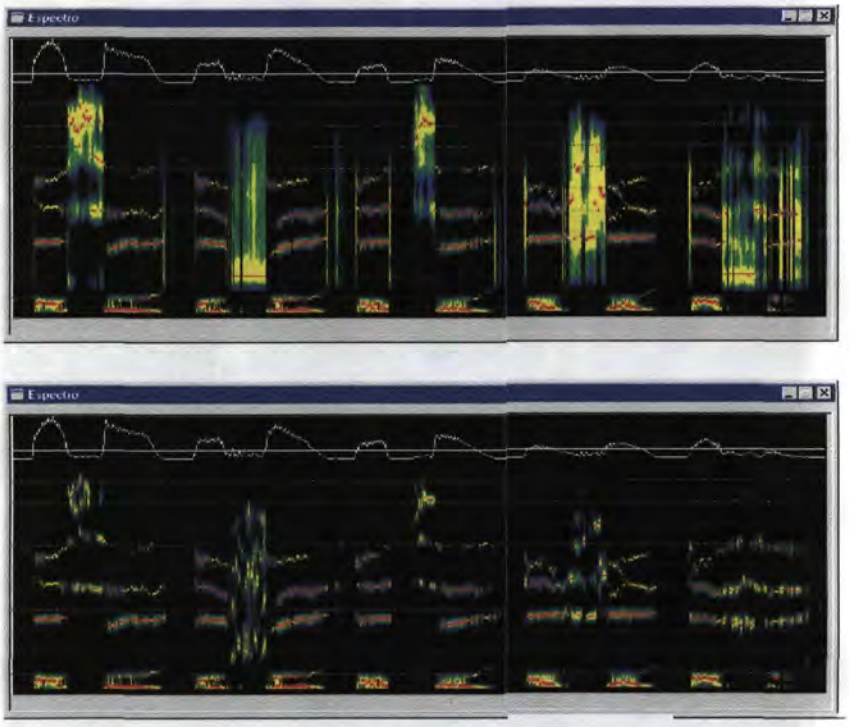


Figura 4.43

Resultado de aplicar un filtro paso bajo a las funciones básicas.

En ambos casos, [f] aparece nítido y fácil de distinguir, siendo los sonidos [x] y [θ] el que más se diferencia entre los dos tipos de espectro. Esto es debido a que sus realizaciones espectrales siguen un modelo más parecido al de los sonidos sonoros, presentándose menos ruido y más estructura de formantes, con lo que en la función de aproximación tienden a confundirse con los sonidos sonoros adyacentes.

Los espectros obtenidos pueden considerarse como característicos, en cuanto a que no se van a realizar modificaciones substanciales sobre los mismos, salvo ajustar algunos parámetros, especialmente la anchura del filtro paso bajo. En la figura 4.44 se presenta el resultado de variar la anchura del filtro paso bajo. Los dos primeros casos se han calculado con la función de polos. El gráfico superior corresponde a un filtrado pequeño y el segundo caso a un filtrado mayor. Los dos últimos espectros han sido calculados con la función de aproximación y filtrados equivalentes.

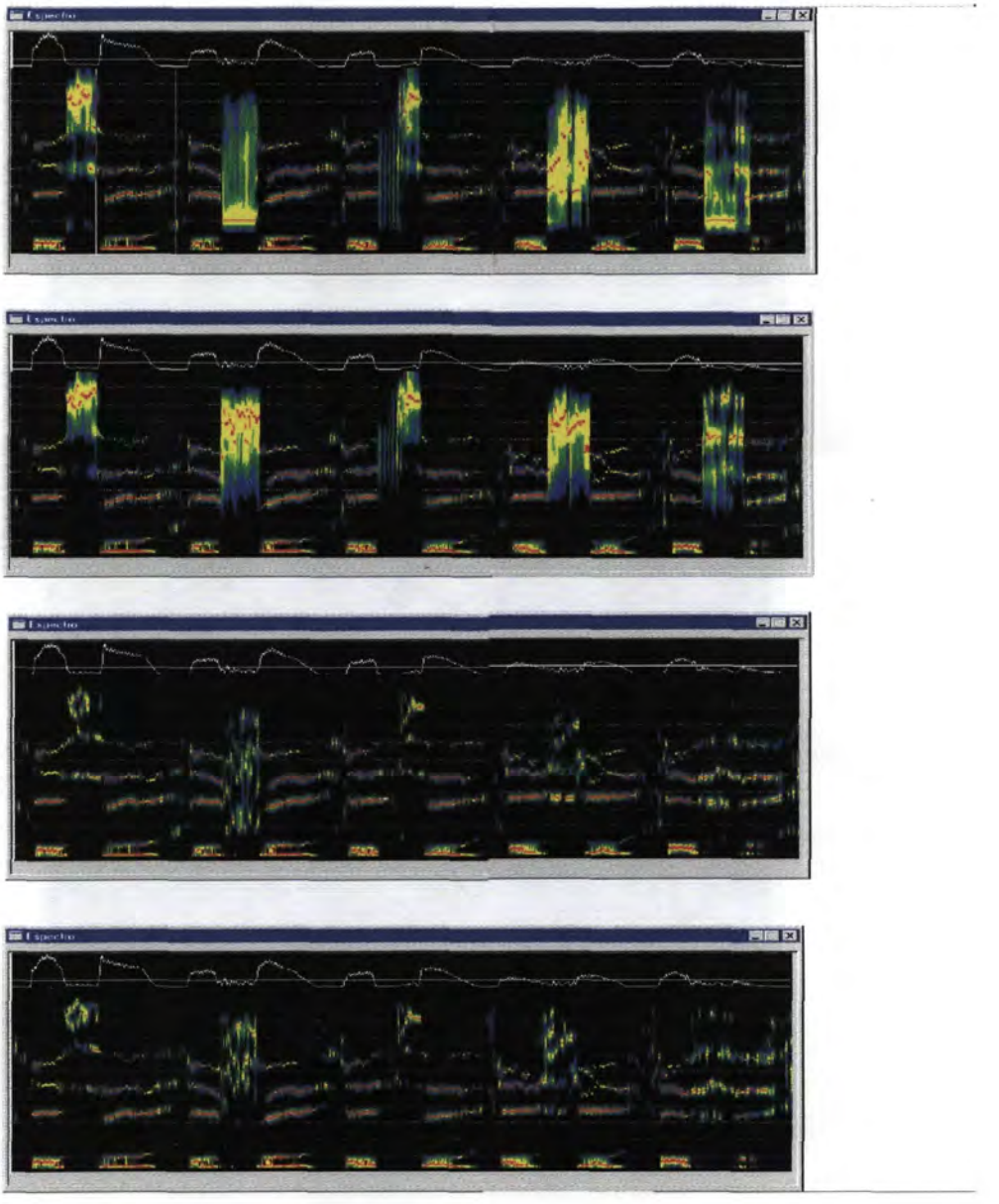


Figura 4.44

Resultados variando la anchura del filtro paso bajo sobre espectros hallados con la función de polos (2 primeros casos) y de aproximación (2 últimos).

Como se puede observar, los resultados obtenidos varían considerablemente de un filtrado a otro, especialmente cuando se usa la función de polos.

Con el fin de estudiar los mismos sonidos sordos rodeados de diferentes vocales, en la figura 4.45 se muestran los resultados correspondientes a las secuencias "oso ofo oθo otʃo oxo", "isi ifi ixi iθi itʃi ixi", "asa afa axa aθa atʃa axa".

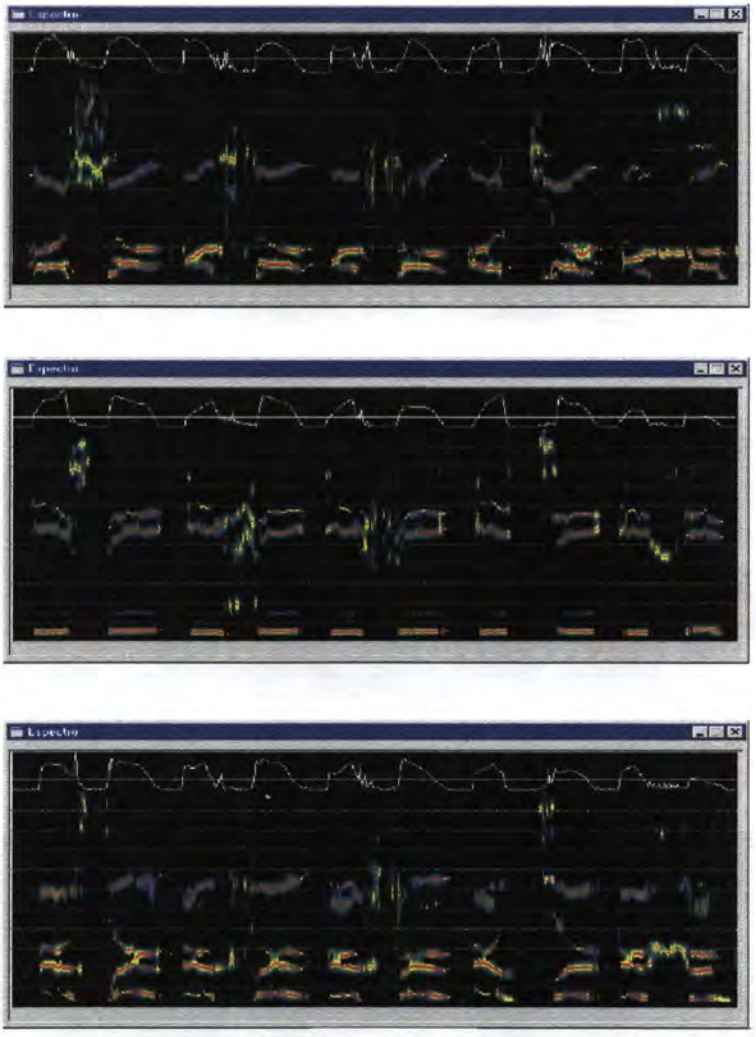


Figura 4.45
Espectros de sonidos sordos articulados con diferentes vocales.

A la vista de estas figuras, se puede determinar que la evolución de los formantes en los sonidos sonoros adyacentes a los sordos es una característica espectral de gran importancia para distinguir estos últimos.

Es adecuado establecer cuidadosamente los umbrales de sensibilidad en la visualización de los espectros para no perder información, tal y como ocurre en la secuencia 'asa' de la figura 4.45.

La figura 4.46 presenta los espectros LPC clásicos equivalentes a los de la figura 4.45. En este caso, se aprecia el problema que existe en la visualización de las frecuencias medias y altas de

los sonidos [f] y [θ], problema que ha sido disminuido en los espectros mejorados mediante la introducción del filtro paso bajo y la normalización frecuencial.

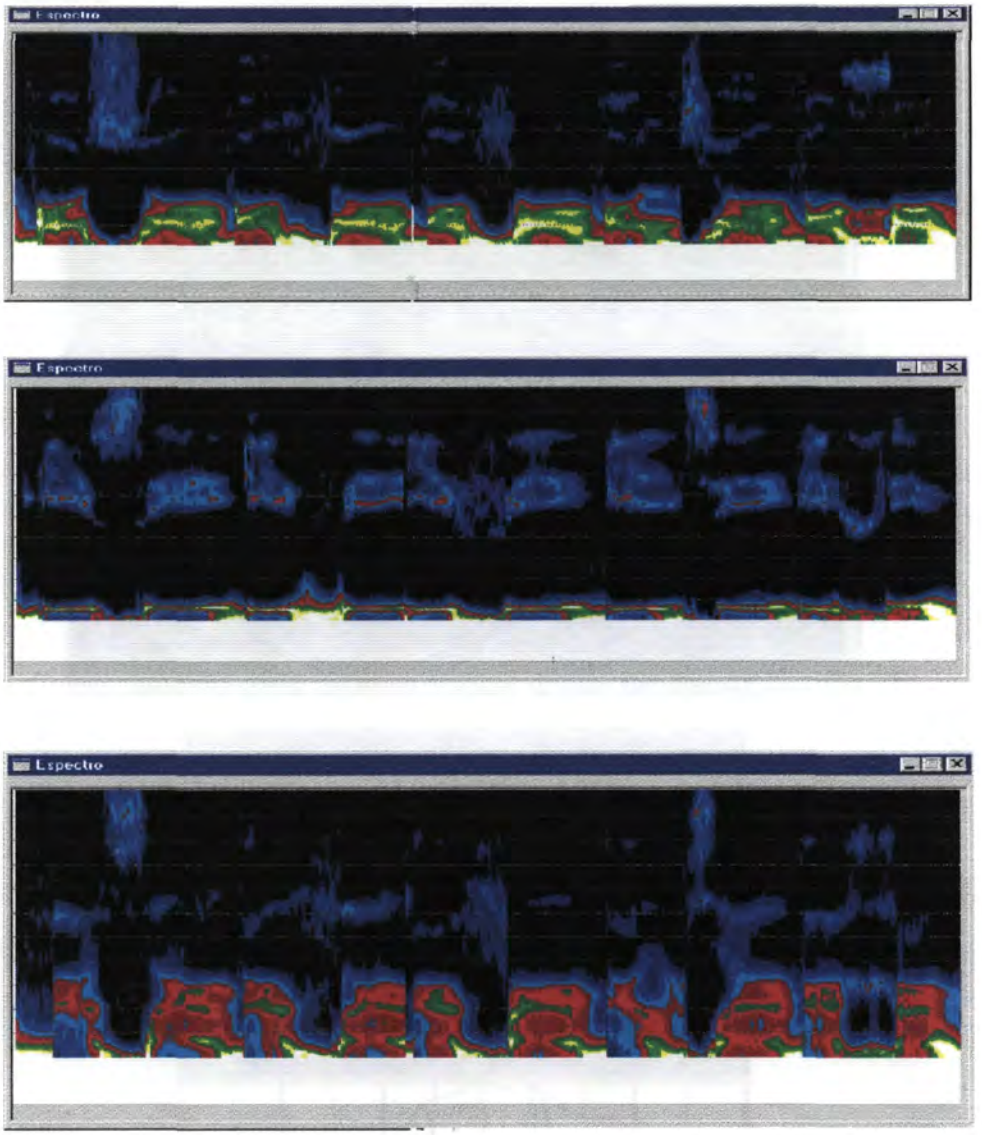


Figura 4.46
Espectros LPC clásicos correspondientes a los de la figura 4.45

En la figura 4.47 se muestra el espectro de las oclusivas sordas en la realización "epe ete eke". Se puede observar la evolución de los formantes hacia un locus bajo en el sonido bilabial, y hacia un locus medio en [t]. El sonido [k] presenta una subida que se manifiesta especialmente en el tercer formante. Todo ello concuerda con el comportamiento esperado.

En cuanto a las barras de oclusión, la de [t] se aprecia con claridad, no siendo así con la de [p] ni la de [k], por ello, en el segundo espectro, se ha empleado la función de polos en las barras

de oclusión, de esta manera, se pueden diferenciar con claridad las distintas zonas (bajas, media, alta) donde se concentra la intensidad en las explosiones de [p], [t] y [k] respectivamente.

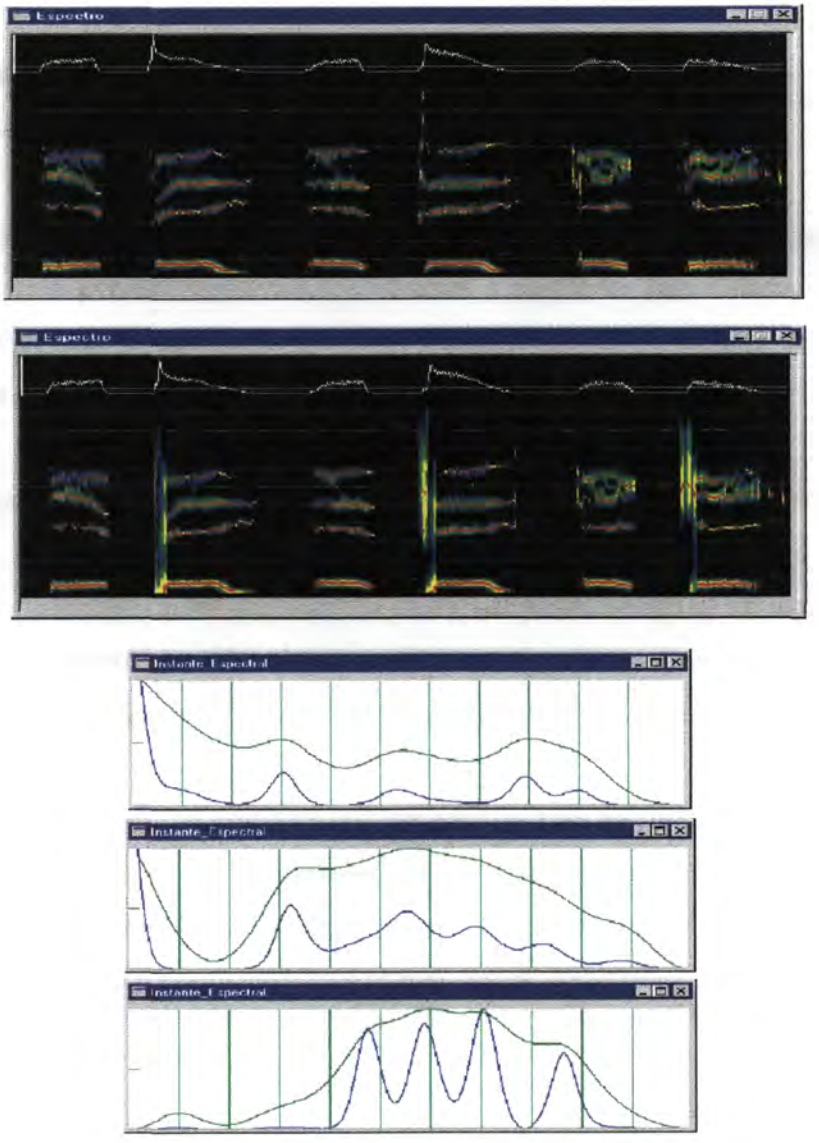


Figura 4.47
Características espectrales de las oclusivas sordas.

Es necesario aclarar que en estos casos de realce con polos, la elección de las zonas de interés se realiza manualmente.

Los tres instantes espectrales de la figura 4.47 nos muestran las características típicas de la función de polos (verde) y de aproximación (azul) correspondientes respectivamente a las explosiones de [p], [t], y [k]. Como se puede observar, mediante la función de aproximación

resulta complicado distinguir [p] de [t], no siendo así con la función de polos, cuya evolución frecuencial varía considerablemente de un caso a otro.

4.4.5 ESPECTROS DE FRASES OBTENIDOS CON LAS FUNCIONES PROPUESTAS

En este apartado se van a mostrar tres frases que sirvan como ejemplos representativos de los resultados que se pueden obtener con las funciones y algoritmos desarrollados. Los espectros conseguidos podrían ser mejorados ajustando manualmente los parámetros de sensibilidad y umbrales de decisión que se han empleado como referencias.

Para cada uno de los ejemplos se ha creado una figura con los siguientes contenidos:

Gráfico a) : Espectro LPC tradicional que se pretende mejorar.

Gráfico b) : Espectro creado empleando la función de aproximación en los sonidos sordos.

Gráfico c) : Espectro obtenido seleccionando manualmente las zonas de sonidos sordos que conviene resaltar y empleando la función de polos en dichas zonas.

La figura 4.48 representa los espectros correspondientes a las palabras "cinco seis siete ocho". En el caso a) (LPC tradicional) se aprecian bastante bien las transiciones entre formantes y el ruido producido por los distintos sonidos [s] y el de [tʃ], que como ya se había estudiado, son los sonidos sordos que mejor se visualizan en todo tipo de espectros.

En el gráfico b) (LPC con función de aproximación), el sonido fricativo con el que comienza la frase aparece con claridad, y también mejora la visualización de los formantes. Por contra, como cabe esperar en este método, las barras de oclusión desaparecen debido a la baja energía que presentan en el hablante que realizó las grabaciones.

El gráfico c) muestra un espectro muy claro y fácil de interpretar. Se puede observar las diferentes realizaciones de las explosiones de las palabras cinco (más alta) y siete (de nivel medio), así como la altura que tiene el ruido de las 'eses' frente al de las demás fricativas que intervienen en la frase. Todo ello acorde a los resultados esperados.

La figura 4.49 ilustra los resultados correspondientes a la frase "os deseo felices fiestas". En el caso LPC tradicional, la sección "os deseo" presenta una calidad adecuada, sin embargo, la parte de la frase "felices fiestas" falla en diversos aspectos, especialmente en que no aparecen las características de los sonidos [f], [l], [t] y [θ].

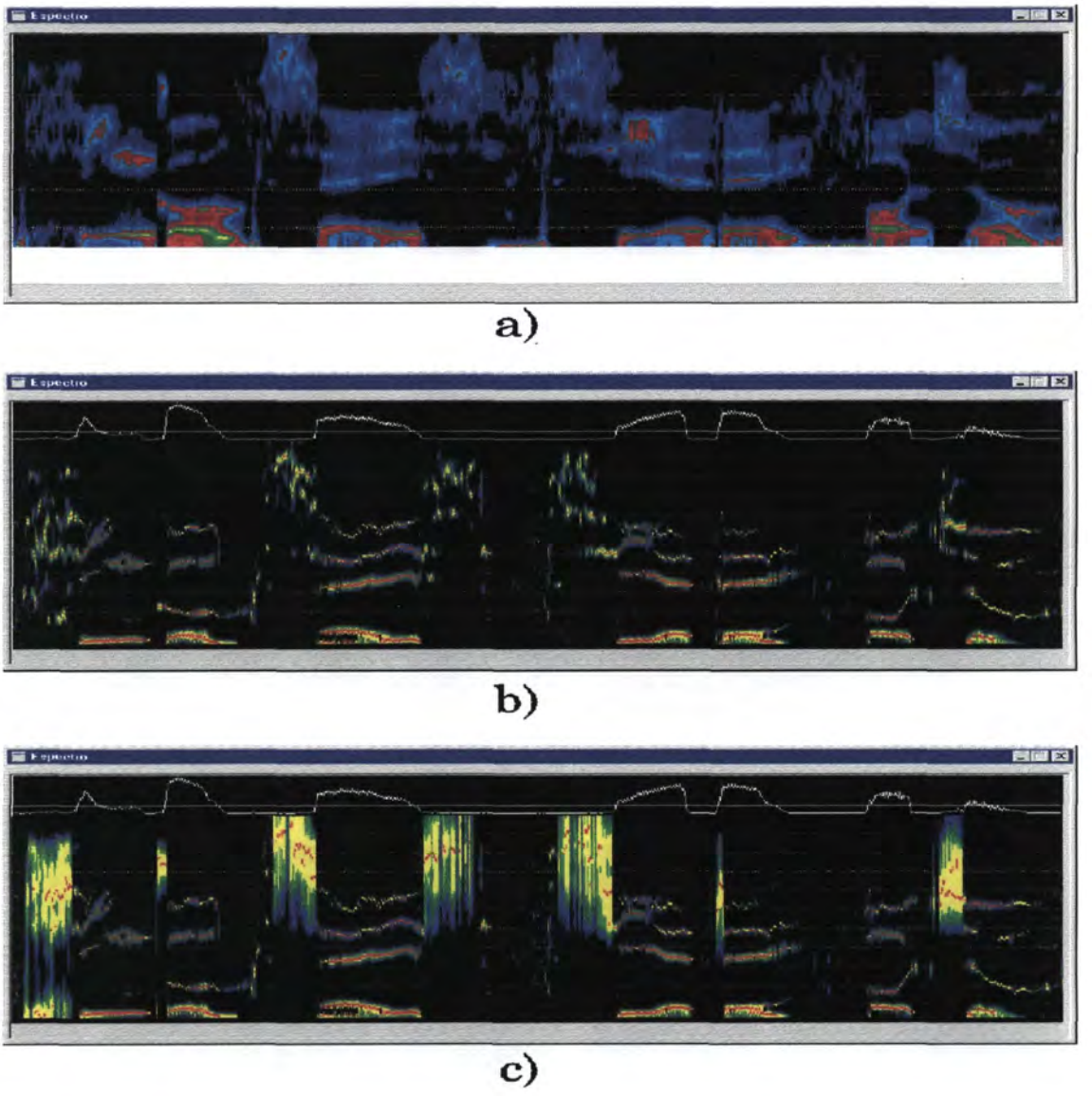


Figura 4.48
Espectros generados a partir de la frase "cinco seis siete ocho".

En el espectro generado con la función de aproximación, los sonidos [f] se recuperan, [l] y [t] también, aunque de nuevo, la explosión de [t] queda sin visualizar. Los formantes, como es habitual, se presentan más claros, continuos y diferenciados.

En el último gráfico de la figura, las 'eses' se distinguen con claridad por la altura de su energía frecuencial, la barra de oclusión aparece gracias a la selección manual, y como único dato negativo, es complicado distinguir los sonidos [f] y [θ].

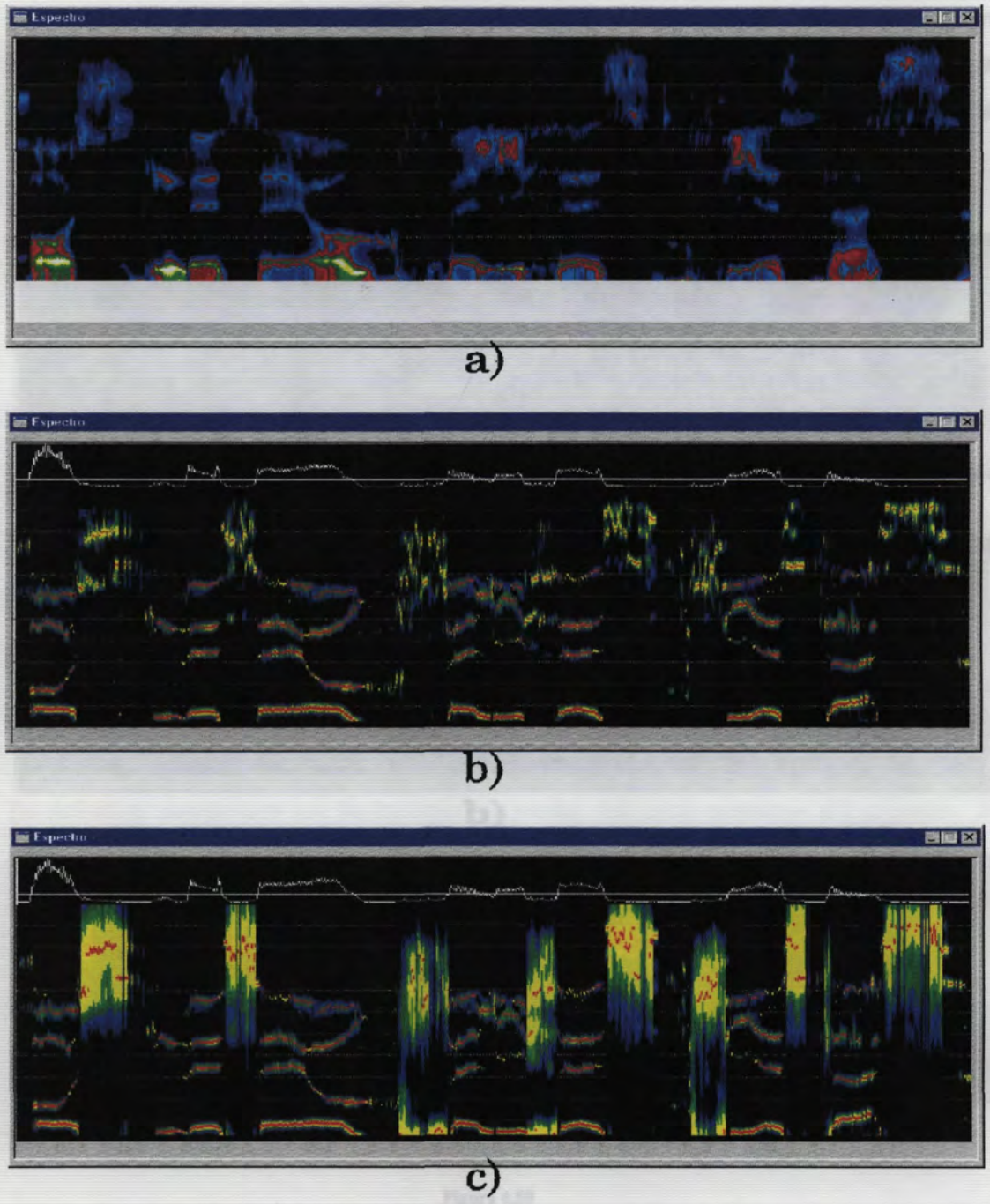
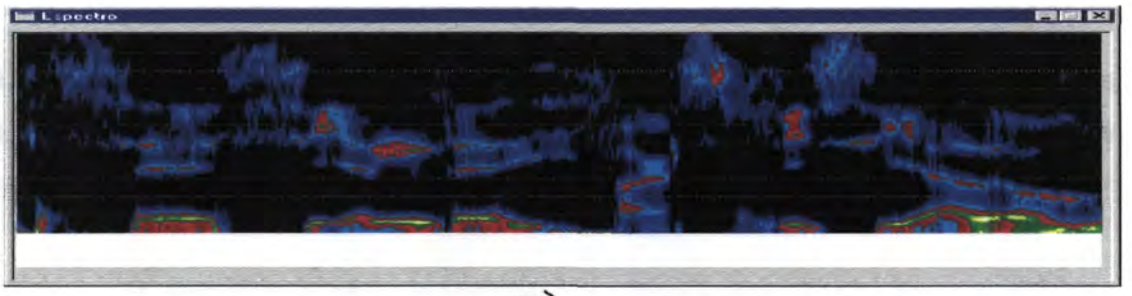


Figura 4.49

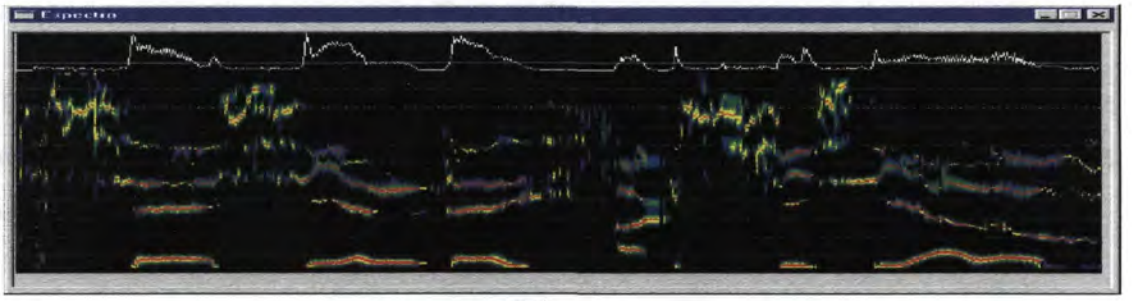
Espectros generados a partir de la frase "os deseo felices fiestas".

La figura 4.50 presenta el ultimo ejemplo a través de una frase que, como las anteriores, ha sido seleccionada por su riqueza en sonidos fricativos, en este caso, la secuencia de estudio es: "se siente asfixiado".

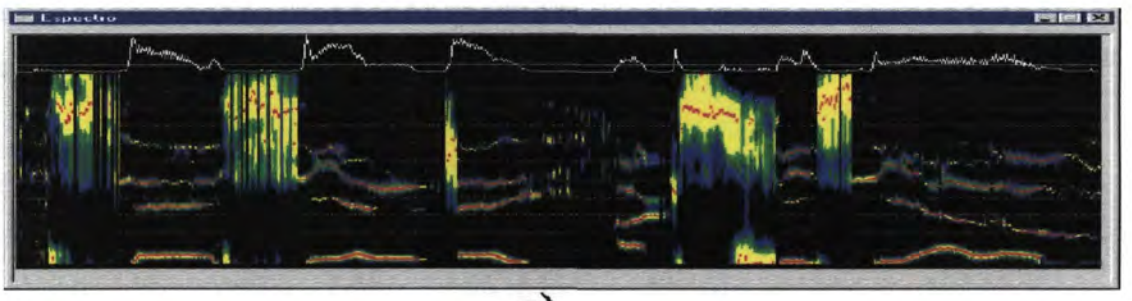
Las características, carencias y mejoras de cada uno de los espectros de la figura, coinciden con los detallados en los ejemplos anteriores. En este caso resulta interesante observar la progresión de los formantes en el final de la frase, y comprobar la dificultad que existe para distinguir el sonido oclusivo sonoro en estos formantes.



a)



b)



c)

Figura 4.50

Espectros generados a partir de la frase "se siente asfixiado".

4.4.6 CONCLUSIONES

Usando la transformación LPC básica, no resulta sencillo visualizar adecuadamente los sonidos del castellano. Esto es debido principalmente a la unión de dos circunstancias:

1.- No se conoce un método que distinga al 100% las características de sordez-sonoridad de una señal de voz continua.

2.- No resulta fácil distinguir sin errores los sonidos sordos de los silencios de las frases.

Una vez constatadas nuestras limitaciones, si se desea ofrecer espectros sin discontinuidades, se ha de optar por una solución conservadora, en la que las características espectrales de los sonidos sordos no aparecen determinadas en su totalidad.

Como opción alternativa, se ofrece la posibilidad de selección manual de las zonas de estudio, y tras ello, la visualización automática de las características espectrales en la zona seleccionada.

También se puede optar por una selección-visualización totalmente automática, que ofrece resultados con discontinuidades, pero con ventajas en la interpretación de los espectros obtenidos.

En definitiva, aunque con sus limitaciones, se pueden conseguir espectros que partiendo de LPC, mejoren considerablemente los resultados obtenidos en los espectros tradicionales en cuanto a la visualización de los sonidos sordos.

4.5 CONCLUSIONES

La conclusión principal que se puede obtener tras el desarrollo de este capítulo es que resulta posible encontrar algoritmos basados en el método de predicción lineal que permiten resaltar características espectrales importantes de la señal de voz. Esta conclusión viene a confirmar la hipótesis sobre la que se ha basado la tesis.

El estudio ha sido enfocado hacia la determinación de los formantes más significativos que se generan en la producción del habla, con ello se hará posible realizar (en el capítulo siguiente) un estudio detallado de la situación y evolución de los formantes en los principales sonidos del español.

La decisión de trabajar con funciones espectrales suavizadas ha resultado muy adecuada para la consecución del fin que se pretendía lograr: la adecuada estimación de los formantes de voz.

Como consecuencia de las pruebas y estudios realizados, se ha considerado necesario hacer uso de diferentes etapas de manipulación de las funciones espectrales básicas, en estas etapas se aplican transformaciones no lineales basadas en métodos algorítmicos.

La determinación de características espectrales en los sonidos sordos se ha realizado sobre la base de que el rasgo de sordez/sonoridad se selecciona manualmente. Asumiendo esta limitación, los métodos ideados contribuyen adecuadamente a la correcta y clara visualización de los sonidos sordos.

Los algoritmos desarrollados proporcionan en su conjunto una buena calidad en la visualización de los espectros de voz, calidad basada en la correcta determinación de las características espectrales buscadas, sin embargo, los tiempos de respuesta no son lo suficientemente cortos y predecibles como para poder ser utilizados como núcleo de aplicaciones de voz con fuertes limitaciones temporales.

5 ANÁLISIS DE LA SITUACIÓN Y EVOLUCIÓN DE FORMANTES EN SONIDOS DE LA LENGUA ESPAÑOLA

5.1 RESUMEN

Una vez desarrollados los métodos de extracción de formantes explicados en el capítulo anterior, los algoritmos obtenidos han sido aplicados a porciones representativas de los sonidos existentes en la lengua castellana. Los resultados se analizan y comparan de forma gráfica y numérica, con el fin de facilitar su comprensión y caracterización.

El estudio que aquí se presenta, pretende ampliar y profundizar los existentes, basados en su mayoría en análisis visuales de espectros básicos, en los que no se puede apreciar los formantes con la suficiente claridad ni precisión.

Los resultados conseguidos mediante técnicas LPC y transformaciones no lineales de señal, al presentarse de forma gráfica, facilitan la caracterización y clasificación de los sonidos en base a la disposición y evolución de sus formantes. Con los valores numéricos obtenidos se realiza un estudio detallado que establece distribuciones, mapas bidimensionales, etc. de los formantes hallados.

5.2 INTRODUCCIÓN

La información más representativa que permite la comprensión de un espectrograma de voz es la posición y evolución de los formantes existentes en los sonidos sonoros [PET52], [KAT95], sin embargo, una vez obtenidos, es necesario conocer sus peculiaridades, que varían

apreciablemente según quien sea el hablante, los distintos contextos que presentan los sonidos, la entonación empleada en las frases, etc. [TOK93]

El presente estudio pretende en primer lugar, establecer las características típicas de los sonidos de la lengua castellana, y tras esta fase clasificadora, presentar las variaciones que de forma natural se producen en el habla.

Con el fin de facilitar la consulta de la información obtenida, se presentan apartados diferenciados según las muestras provengan de uno o varios hablantes. También se establecen apartados distintos atendiendo a la naturaleza de los datos: sonidos vocálicos o no vocálicos. Por último, se remarcan las ideas principales en un apartado de conclusiones.

5.3 SONIDOS VOCÁLICOS EN UN SÓLO HABLANTE

5.3.1 INTRODUCCIÓN

Antes de intentar generalizar y formalizar los resultados obtenidos en el estudio de los formantes de la lengua castellana, resulta adecuado comenzar con el caso más sencillo, de forma que se puedan ir estableciendo conclusiones con un sólo hablante que sirvan por una parte como un caso de estudio válido, y por otra como referencia sobre la que se añadirán futuros resultados.

En primer lugar se realizará un estudio completo basado en un sólo hablante. Como comienzo se abordan los sonidos vocálicos y se explica la forma en que se presentan los resultados.

Resulta necesario establecer las porciones de sonidos que van a ser estudiadas. Un enfoque tradicional para su elección, se basa en la selección de los sonidos básicos y los alófonos más importantes que presenta el español, todo ello accesible mediante la bibliografía clásica de fonética española [QUI93], [MAR94]. En lugar de ello, se ha considerado más oportuno realizar grabaciones de los sonidos básicos rodeados de diferentes contextos, por ejemplo: 'epe', 'apa', 'ete', 'usu', etc. Con ello, se evita el enfoque articulatorio que impera en las clasificaciones típicas, buscándose las variaciones de los sonidos desde un estudio basado totalmente en la fonética acústica.

Los sonidos han sido grabados atendiendo al modo de articulación, con el fin de poder realizar comparaciones de los resultados basadas en esta clasificación. Los grupos elegidos son:

- Vocales aisladas: 'i e a o u'
- Nasales: 'imi ini iɲi', 'eme ene eɲe', 'ama ana aɲa', 'omo ono oɲo', 'umu unu uɲu'
- Oclusivas sordas: 'ipi iti iki', ..., 'upu utu uku'
- Fricativas sonoras: 'iβi id.i iɣi', ..., 'uβu ud.u uɣu'
- Fricativas y africada: 'ifi iθi isi ixi itʃi', ..., 'ufu uθu usu uxu utʃu'
- Laterales-Vibrantes: 'ili iri iʎi irri', ..., 'ulu uru uʎu urru'

En este apartado se presentan los espectros de los sonidos seleccionados, y además gráficos de la situación de los tres primeros formantes de las vocales castellanas, tanto de forma aislada como en la proximidad de cada grupo consonántico.

Los espectros de voz y la estimación de formantes se ha realizado aplicando los métodos y algoritmos ideados en el capítulo anterior.

Aunque en la bibliografía tradicional habitualmente se compara únicamente la evolución de los dos primeros formantes [PET52], [HIL95], [MAR94], en este caso, confrontaremos los tres primeros, tomados de dos en dos en sus tres posibles combinaciones.

5.3.2. DESARROLLO

Para comenzar, se presenta en la figura 5.1 el espectro de la voz obtenido a partir de una secuencia de sonidos 'i e a o u'. Como se puede observar, los formantes de las vocales se aprecian visualmente con claridad, salvo F2 en [i], que no aparece.

Las cuatro gráficas inferiores de la figura representan instantes espectrales significativos de las vocales 'i', 'e', 'o' y 'u'. Tal y como se explicó en el capítulo anterior, los formantes se obtienen a partir de estas funciones. Es importante resaltar que en el caso de la 'i', se detectan los tres formantes (gráfico superior izquierdo), a pesar de no ser visualizados en su totalidad.

La figura 5.2 presenta un gráfico tridimensional (imagen superior izquierda) en el que cada eje representa un formante (X:F2, Y:F3, Z:F1). Cada superficie de la figura está formada por las

posiciones de los tres primeros formantes de distintos instantes espectrales correspondientes a una misma vocal. Todos estos puntos tridimensionales (F2, F3, F1) forman los vértices de la superficie.

Las vocales se identifican fácilmente por las posiciones bien conocidas de F1 y F2. Estudiando el caso, por ejemplo con F2, sabemos que la posición más alta corresponde a la 'i' (rojo), después 'e' (verde), 'a' (azul claro), 'o' (azul), y por último 'u' (morado).

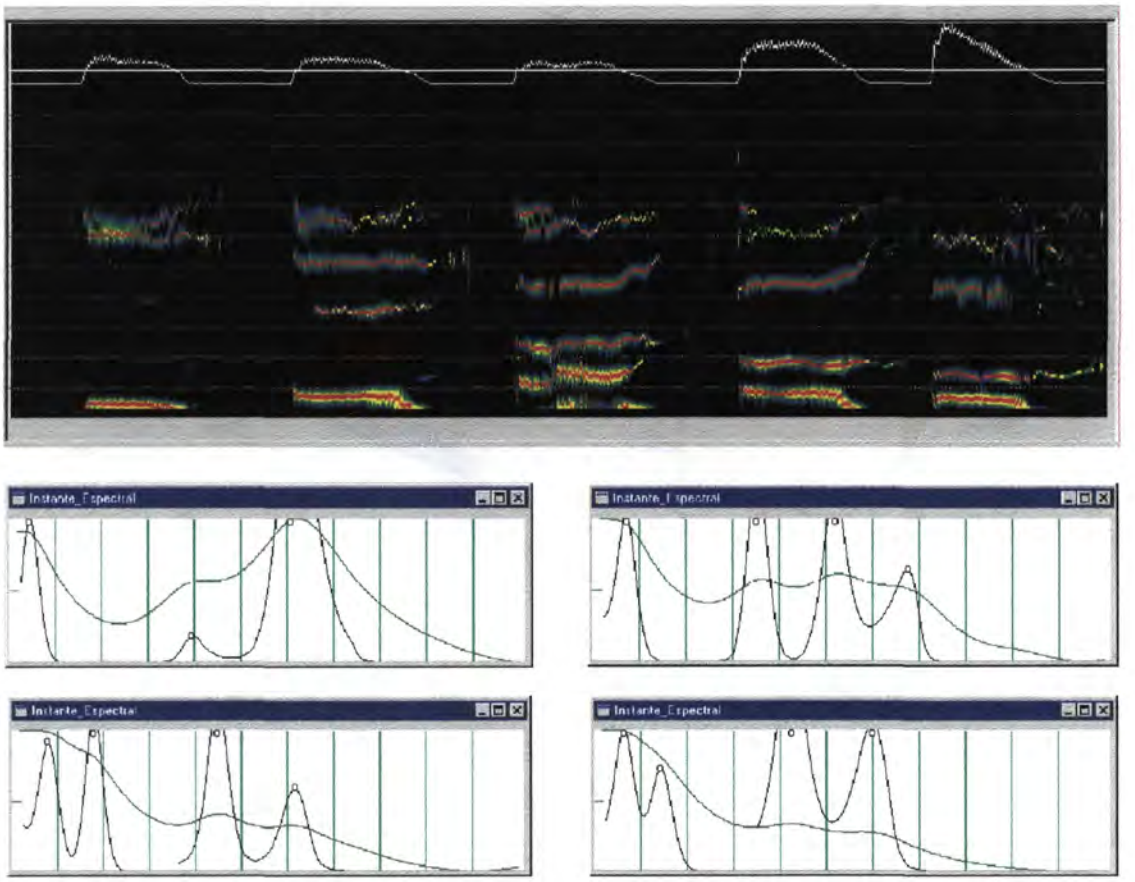


Figura 5.1
Espectro de voz de la secuencia 'i e a o u', junto a varios instantes espectrales de referencia.

Los otros tres gráficos de la figura, representan las posibles proyecciones en el plano de los valores tridimensionales, con esto conseguimos conocer las posiciones de los tres formantes comparados entre sí.

En este caso, se observa que existe una clara separabilidad de las vocales atendiendo a cualquier combinación de formantes. También se aprecia una mayor variabilidad del primer formante de la 'a'. Estas indicaciones se dan a modo de ejemplo, puesto que en este caso se utilizan unas pocas muestras de un sólo hablante, con lo que no se deben generalizar los resultados.

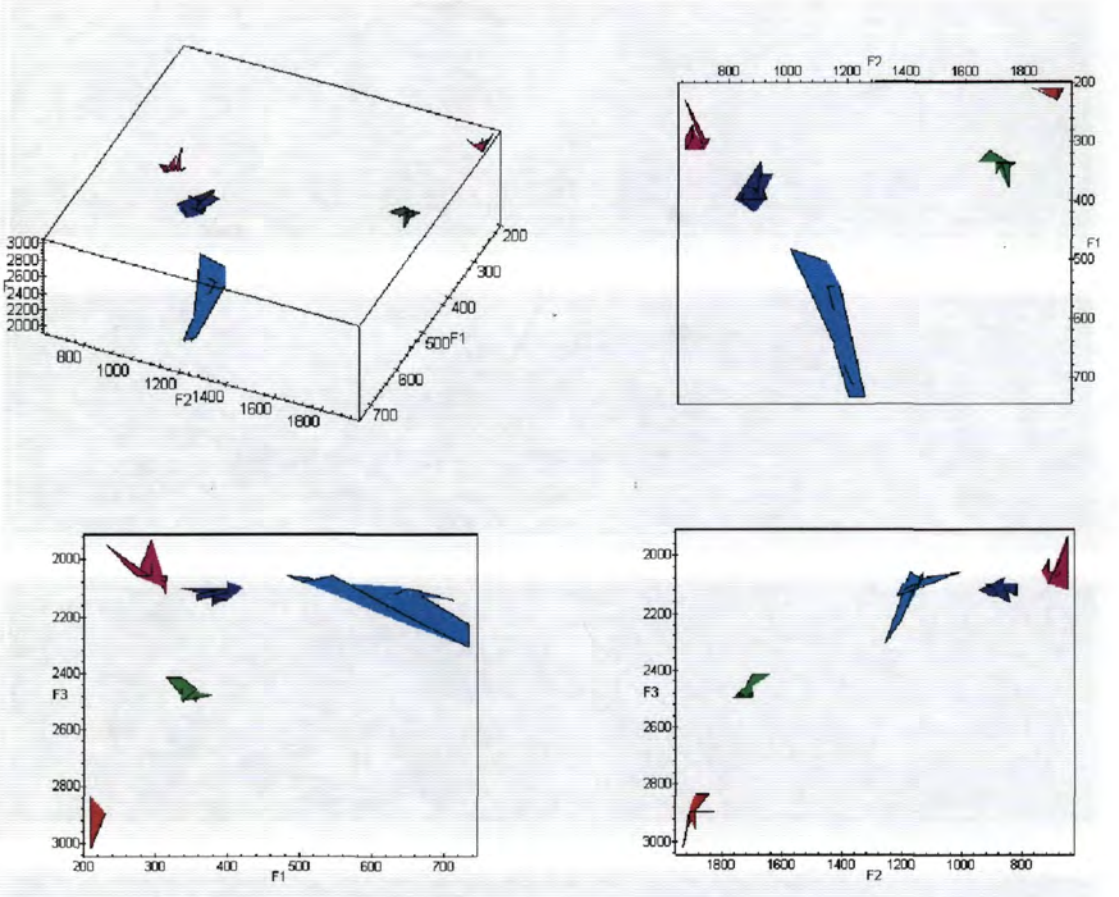


Figura 5.2
Posiciones de los tres primeros formantes de vocales aisladas confrontadas entre sí.

En la figura 5.3 se muestran los espectros obtenidos a partir de las grabaciones correspondientes al grupo nasal. El espectro superior corresponde a los sonidos 'imi ini iɳi', el siguiente a 'eme ene eɳe', y así sucesivamente con 'a', 'o', 'u'.

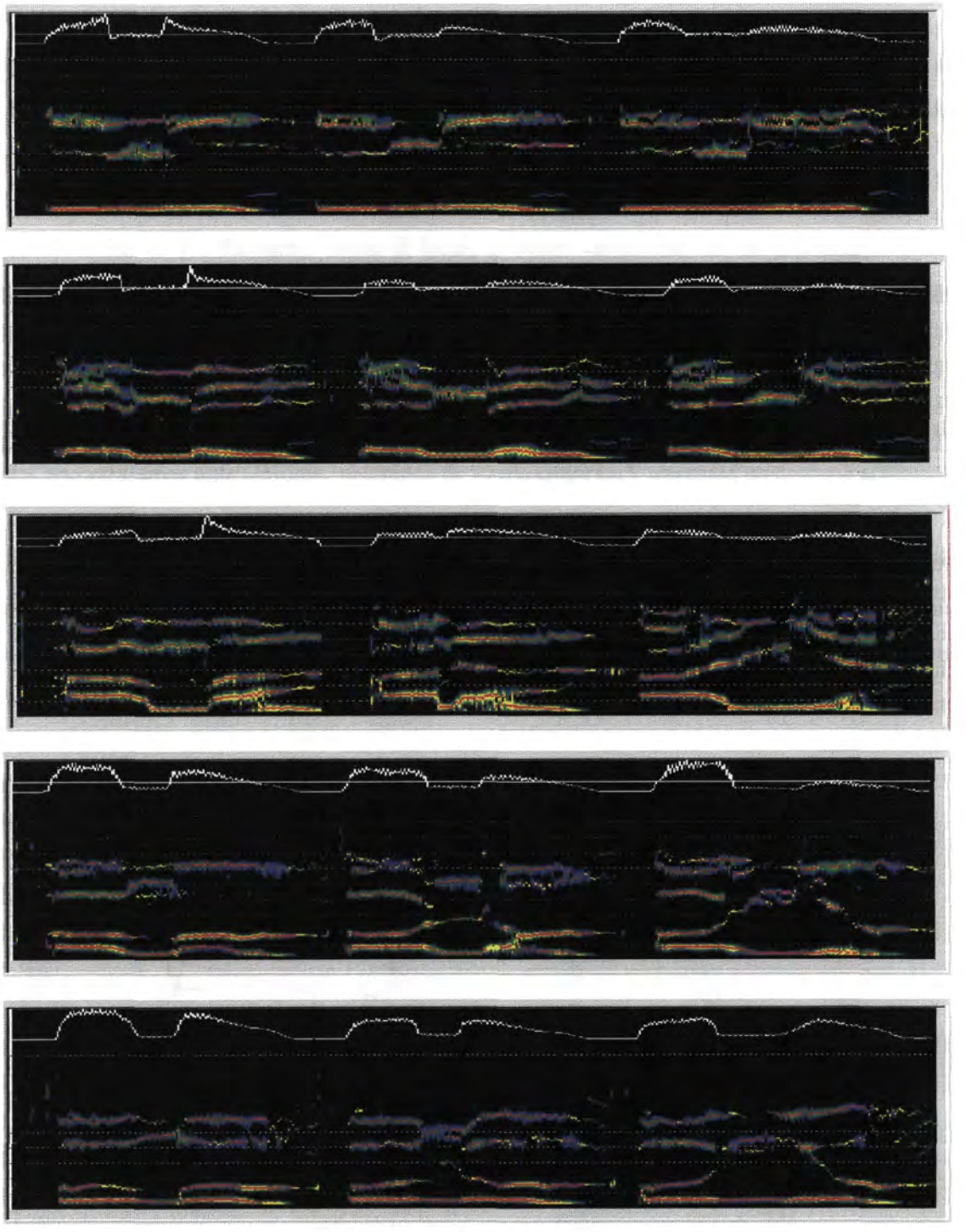


Figura 5.3

Espectros de voz correspondientes al grupo nasal ('imi ini iɳi', 'eme ene eɳe', 'ama ana aɳa', 'omo ono oɳo', 'umu unu uɳu').

Como se puede apreciar, las consonantes nasales presentan unos formantes característicos, pero esto será analizado en otros apartados, centrándonos ahora en sus vocales adyacentes.

Como es sabido, los formantes de las vocales adyacentes a algunas consonantes, tienden a evolucionar hacia el locus correspondiente. En el caso de [m] (bilabial) el locus es bajo, en el de [n] (interdental-dental) medio, y en el de la [ɳ] (velar) alto. Esta situación se repite con los sonidos oclusivos sordos [p], [t], [k] y fricativos sonoros [β], [d.], [γ].

La evolución de los formantes hacia el locus, produce variaciones en las posiciones de los formantes de las vocales, por ello, estudiaremos también estas posiciones, que enriquecerán los resultados que se presentan gráficamente.

Cuando se encuentren variaciones muy significativas respecto a la totalidad de los resultados, nos encontraremos ante alófonos, obtenidos de forma empírica y mediante un enfoque acústico.

La figura 5.4 está formada a partir de la figura 5.3, ampliándose con los formantes obtenidos al evaluar las nuevas muestras. En color amarillo aparecen las superficies ya explicadas, las demás superficies se corresponden a las cinco vocales precedidas y seguidas de consonantes nasales.

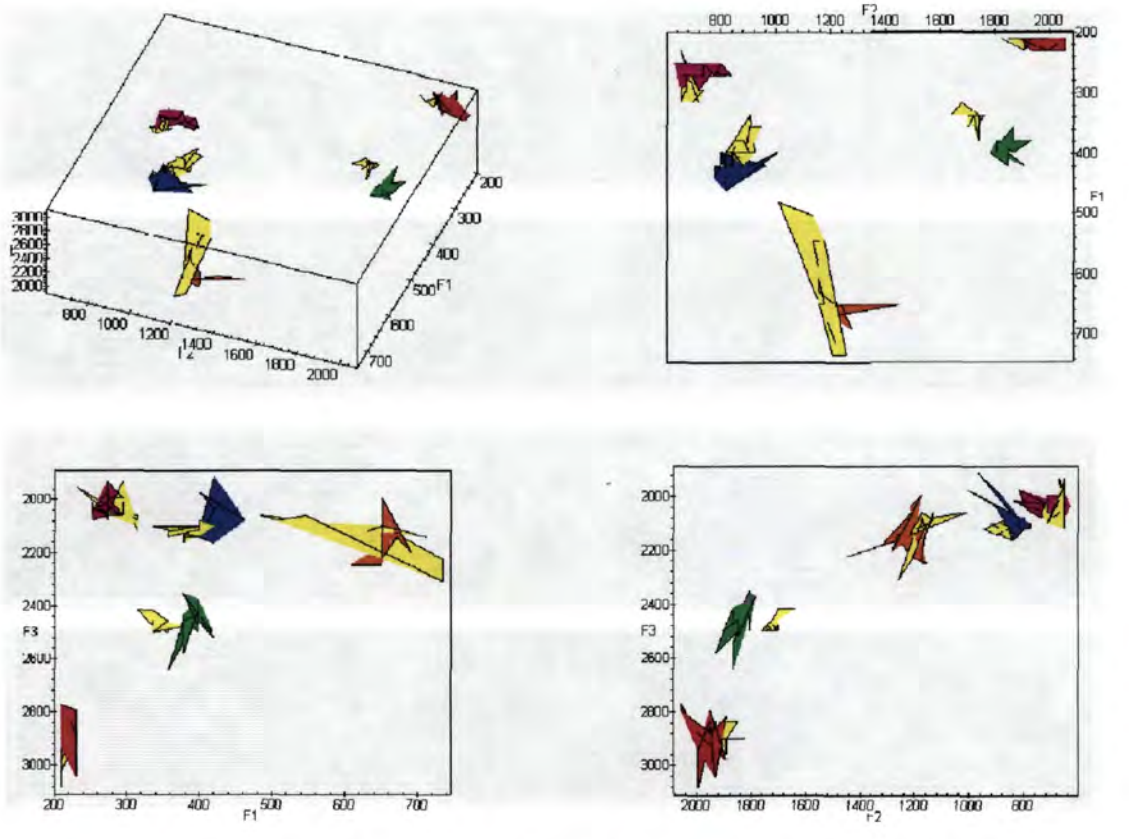


Figura 5.4

Posiciones de los tres primeros formantes de vocales ante nasal confrontadas entre sí.

Como se puede observar, aunque los resultados siguen la misma pauta que los de la figura 5.3, existen claras diferencias entre vocales aisladas y vocales ante nasal, lo que confirma la existencia de los alófonos nasales de las vocales del castellano.

Como ejemplos claros de las diferencias mencionadas, se pueden observar los casos de la 'e' y la 'o' en la representación F1-F2.

En la figura 5.5 se representan los espectros de voz correspondientes al grupo oclusivo sordo, al igual que en el caso anterior, cada formante evoluciona hacia su locus correspondiente, esto se aprecia con gran claridad en este ejemplo en el sonido 'epe' y en el 'aka'.

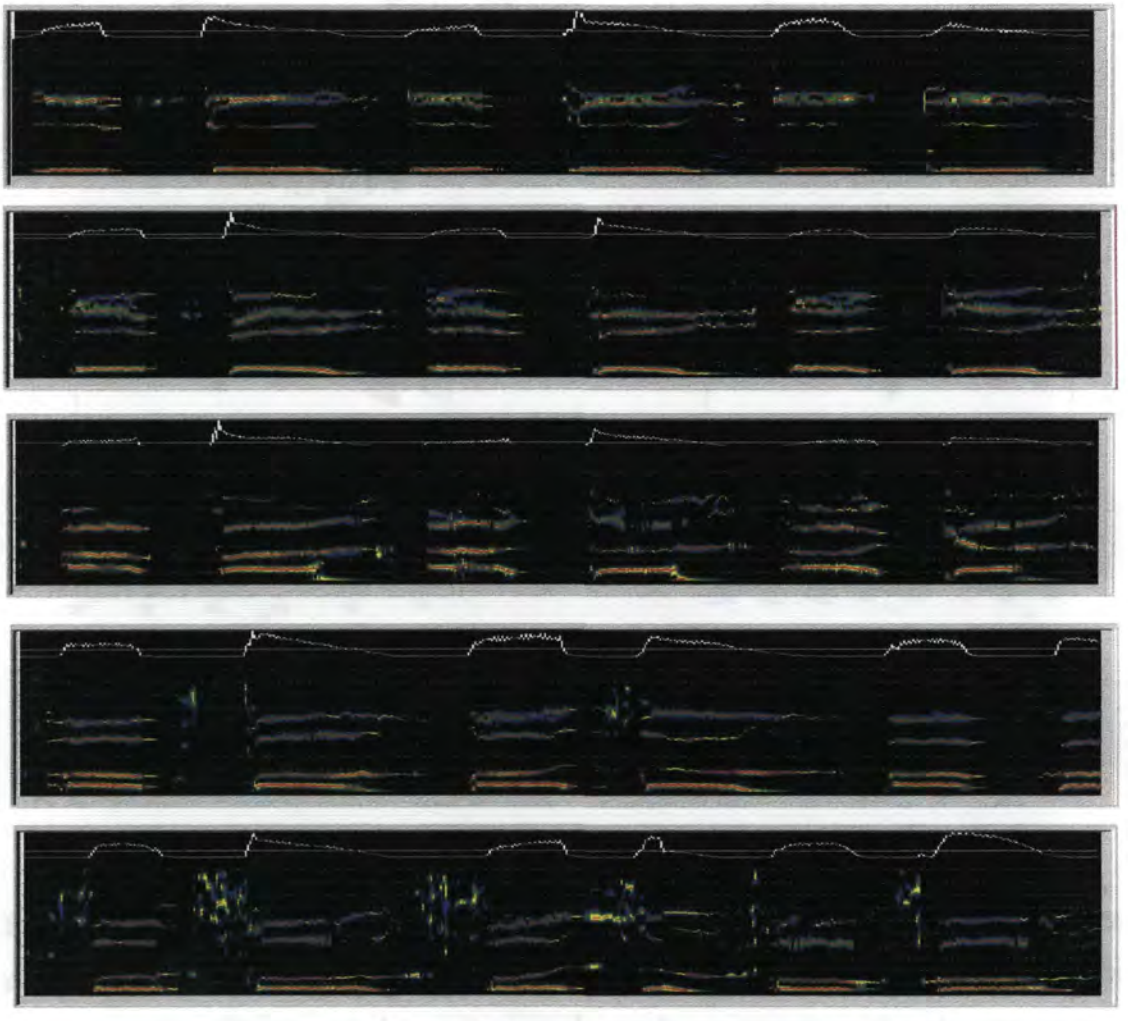


Figura 5.5

Espectros de voz correspondientes al grupo oclusivo sordo ('ipi iti iki', 'epe ete eke', 'apa ata aka', 'opo oto oko', 'upu utu uku').

En los espectros de la figura, las barras de oclusión no aparecen debido a que han sido filtradas junto a los silencios, con el fin de resaltar los formantes de las vocales que son el objeto de nuestro estudio.

Los resultados obtenidos se representan en la figura 5.6. La característica más remarcable es el parecido existente entre los datos obtenidos en las vocales con nasales (figura 5.4) y el caso que nos ocupa. Este parecido se refleja en la disposición de las superficies halladas respecto a las obtenidas en vocales aisladas (en color amarillo).

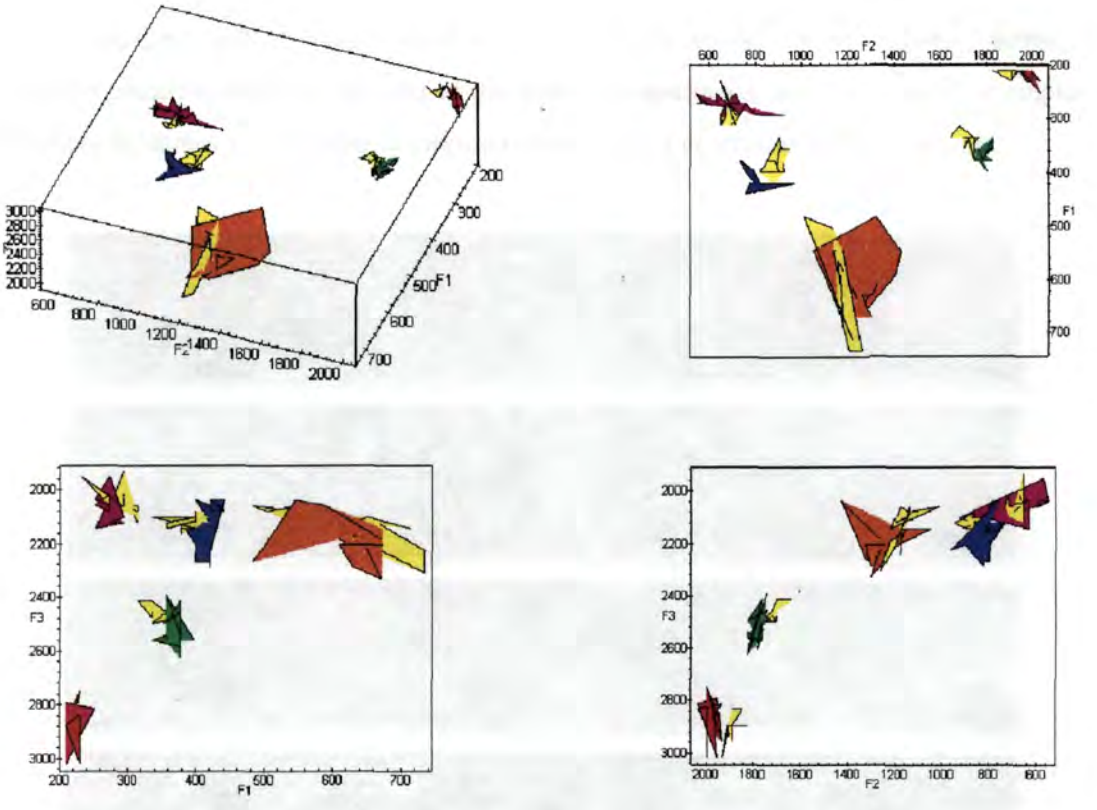


Figura 5.6
Posiciones de los tres primeros formantes de vocales junto a sonidos oclusivos sordos confrontados entre sí.

En la figura 5.7 se presenta el grupo fricativo sonoro. A simple vista se aprecia un enorme parecido en la situación y evolución de los formantes respecto al grupo de las consonantes oclusivas sordas. Obviamente el espacio ocupado por la consonante en sí varía, pero en este apartado se estudian únicamente las vocales cercanas.

En la figura 5.8 aparecen los resultados en los diagramas bidimensionales y tridimensionales habituales. Se puede apreciar una gran similitud entre los grupos oclusivo sordo por un lado y fricativo sonoro por el otro, que lleva a la conclusión de que los formantes de las vocales adyacentes a los sonidos consonánticos estudiados no nos ayudan a distinguir con claridad [p] de [β], [t] de [d.], ni [k] de [γ] [MAR94], [BLU79], [SMI94]. Dicho de otra forma, la posición de los locus en los grupos oclusivos depende del punto de articulación, pero no del modo de articulación. Este razonamiento puede ser extendido (con un mayor margen de variación) al grupo nasal.

Las conclusiones que se van obteniendo a raíz de los resultados presentados, deben ser valoradas como hipótesis a confirmar en los siguientes apartados, donde el estudio se amplía a diferentes hablantes y se utilizan métodos estadísticos para contrastar dichas hipótesis.

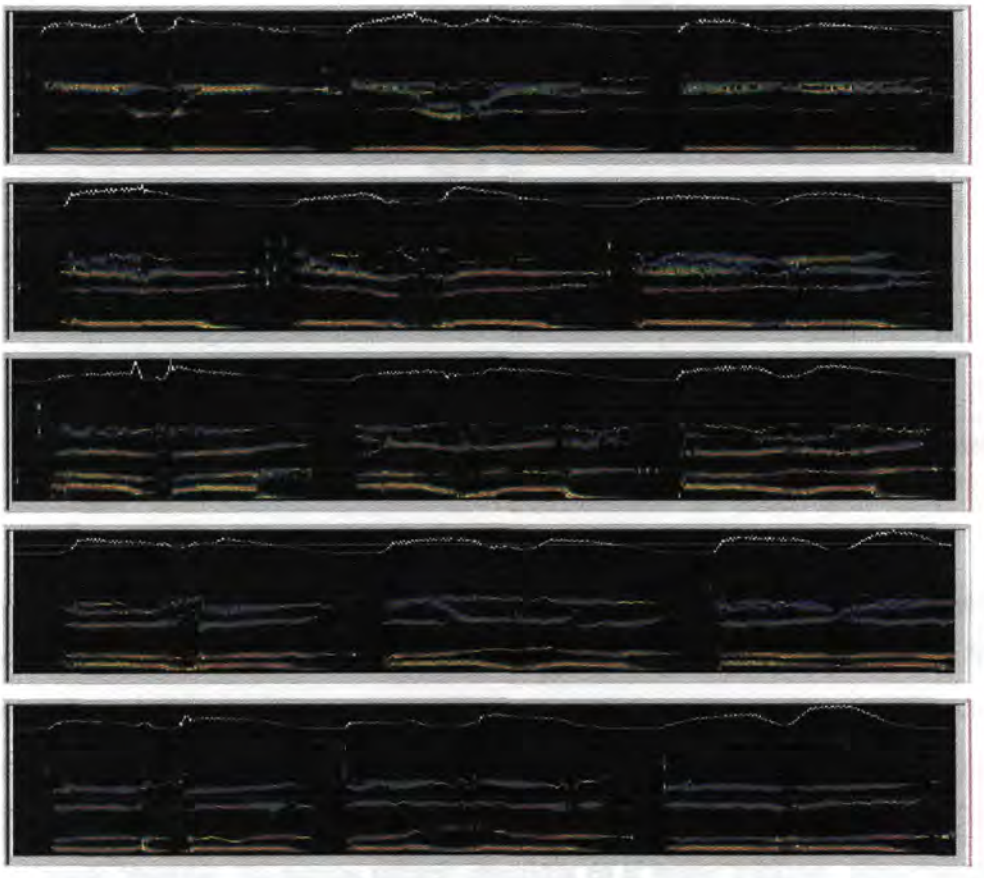


Figura 5.7

Espectros de voz correspondientes al grupo fricativo sonoro ('iβi id.i iγi', 'eβe ed.e eγe', 'aβa ad.a aγa', 'oβo od.o oγo', 'uβu ud.u uγu').

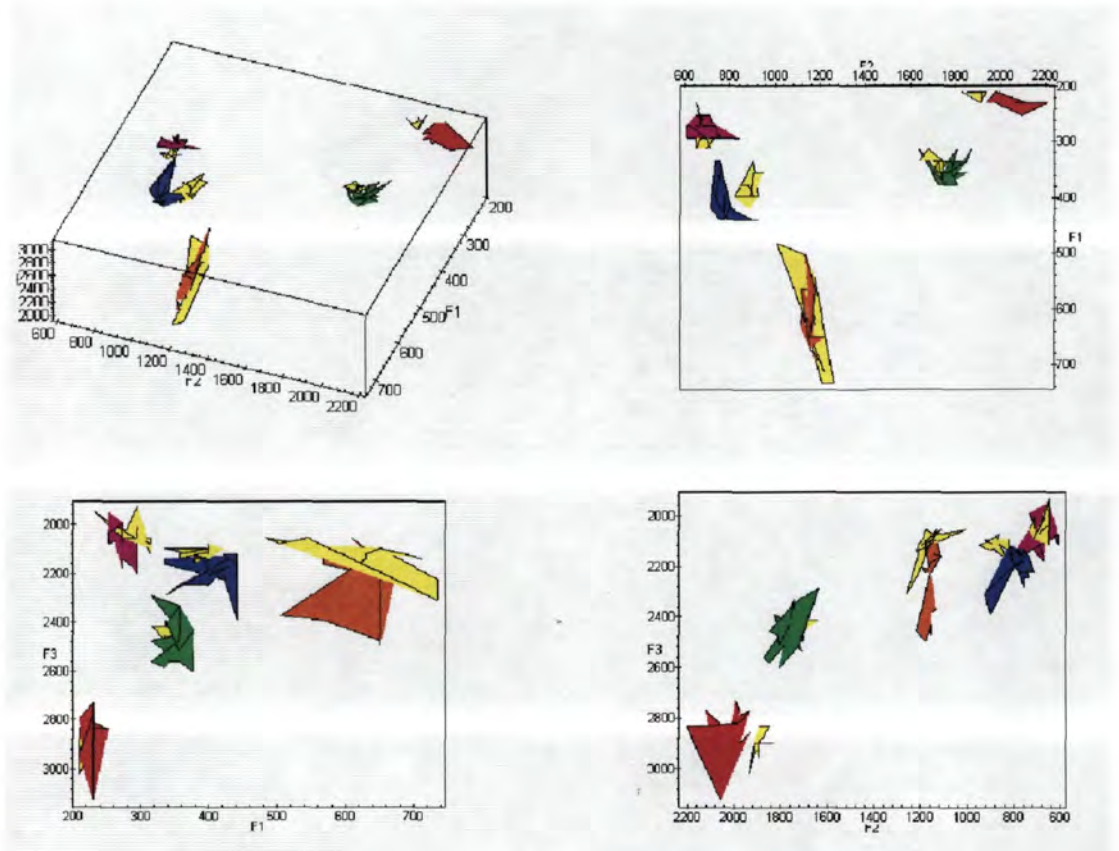


Figura 5.8

Posiciones de los tres primeros formantes de vocales junto a sonidos fricativos sonoros confrontados entre sí.

La figura 5.9 representa a los sonidos fricativos y africado siguientes [f], [θ], [s], [x], [tʃ]. El ruido característico de estos sonidos ha sido filtrado para poder resaltar los formantes con mayor facilidad.

A la vista de la figura 5.10, podemos establecer una mayor altura del primer formante en la 'i' y la 'o' respecto a las mismas vocales aisladas. La misma situación de da con F3 en la 'i' y la 'e'. Estos resultados, aunque son muy parciales para poder establecer reglas seguras, denotan una altura apreciable (que tiende a elevar los formantes) en los sonidos [s] y especialmente [tʃ].

Como se puede observar, los nuevos valores obtenidos en el grupo fricativo, dificultan algo la separabilidad automática de las vocales.

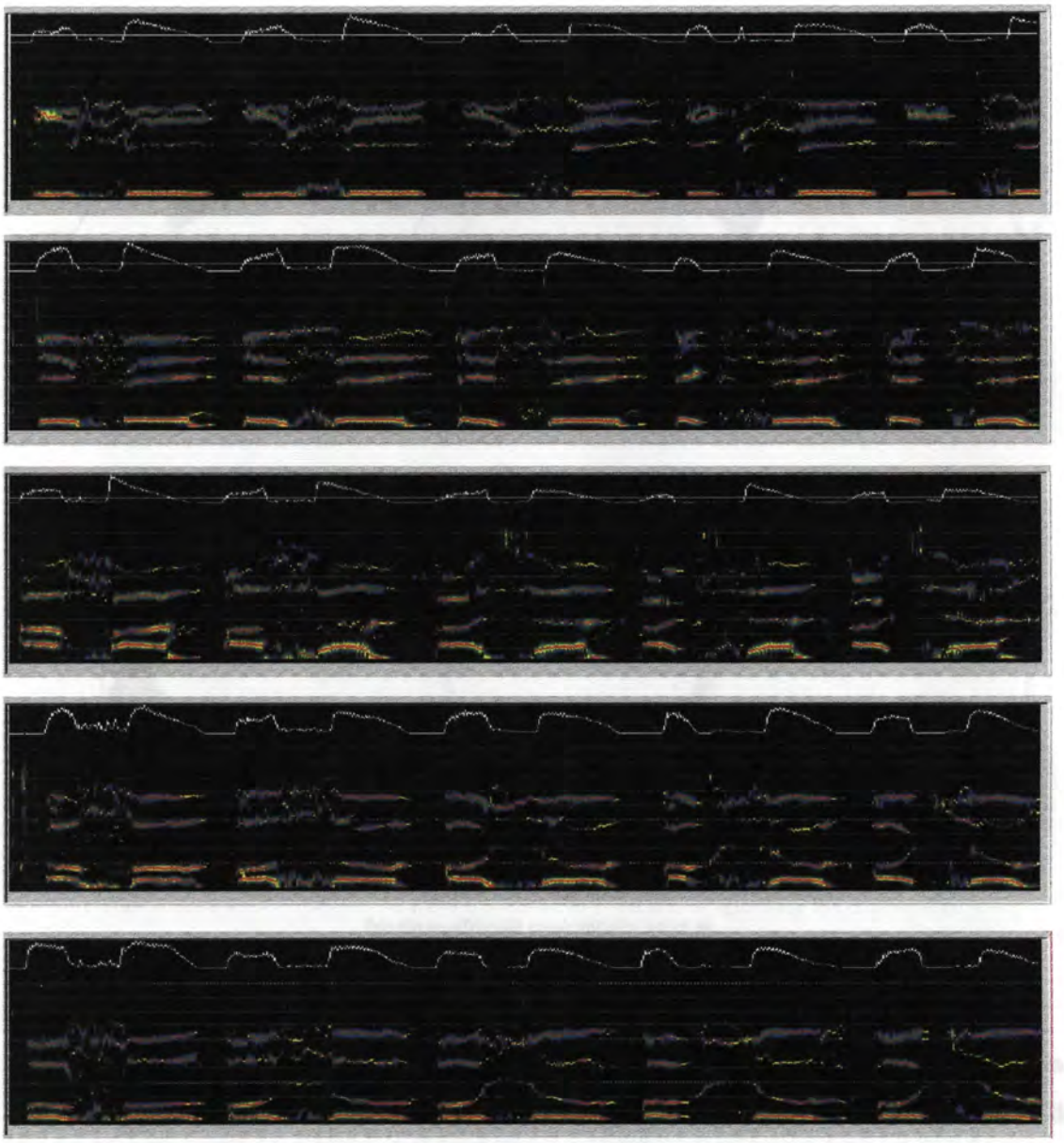


Figura 5.9

Espectros de voz correspondientes al grupo fricativo-africada ('ifi iθi isi ixi itʃi', 'efe eθe ese exe etʃe', 'afa aθa asa axa atʃa', 'ofo oθo oso oxo otʃo', 'ufu uθu usu uxu utʃu').

En la figura 5.11 se presentan algunos espectros de voz del grupo laterales-vibrantes con la secuencia [l], [r], [λ], [rr]. En la figura 5.12 aparecen los resultados de la forma habitual, no se aprecian características especiales que diferencien este grupo de los demás, atendiendo a los parámetros estudiados.

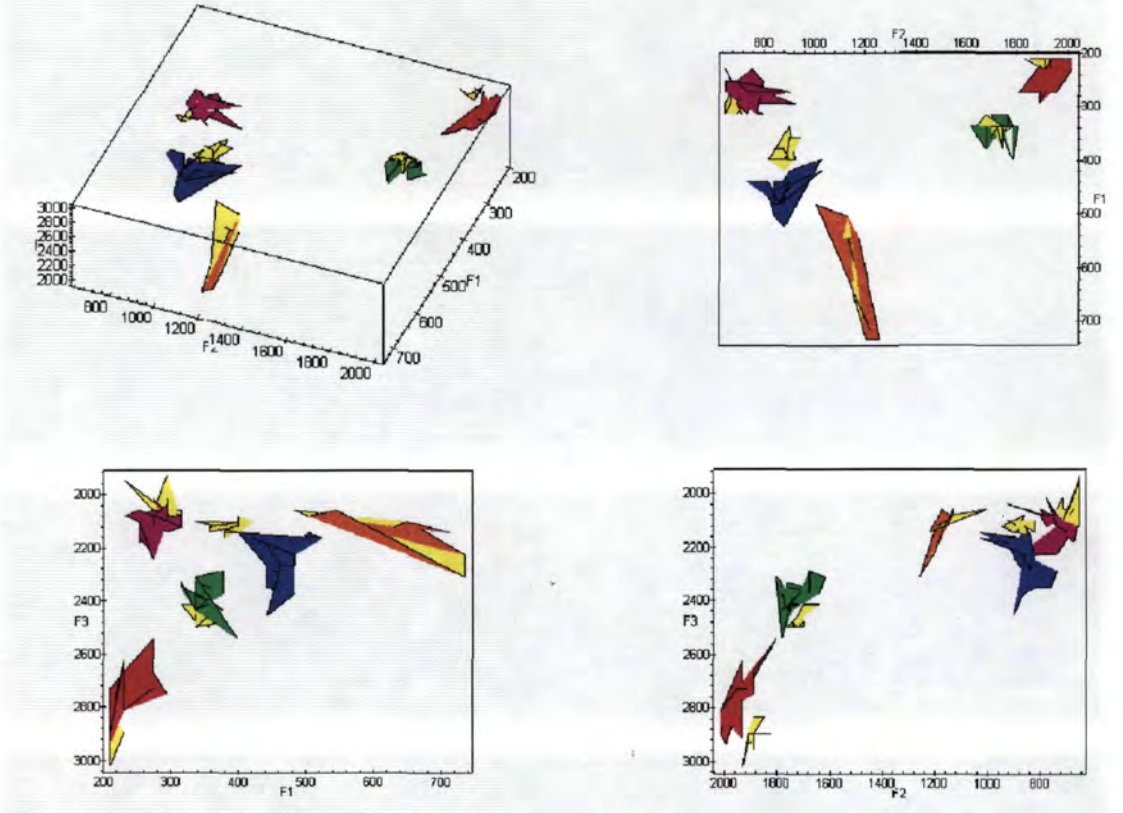


Figura 5.10

Posiciones de los tres primeros formantes de vocales junto a sonidos fricativos-africado confrontados entre sí.

En la figura 5.13 se han colocado todos los grupos estudiados, con el fin de poder realizar comparaciones. Las vocales aisladas se representan de color amarillo, y nos siguen sirviendo de referencia base. En el sonido 'a' se presenta una gran variabilidad en el rango que puede tomar F1, esto también ocurre con el grupo de las oclusivas sordas (verde) y fricativas sonoras (plateado).

El grupo nasal (rojo) se solapa con todas las oclusivas, como ya se observó previamente. Las fricativas (azul), difieren ligeramente en el rango de frecuencias de los tres formantes en los que se representan, comparadas con el resto de los grupos vocálicos.

La figura 5.14 representa las cinco vocales estudiadas, sin realizar diferenciaciones según el origen de las vocales de donde se han tomado las muestras.

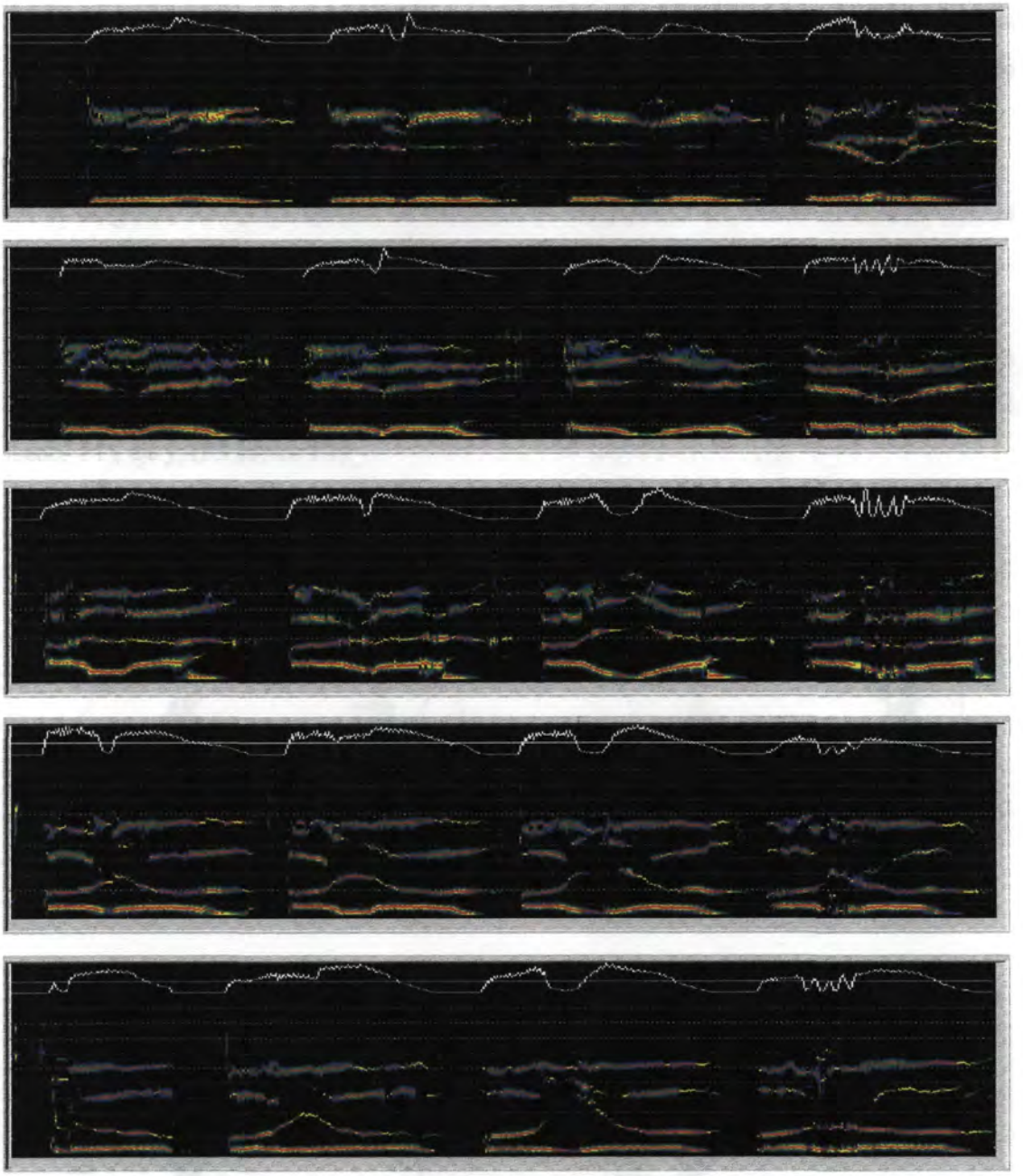


Figura 5.11

Espectros de voz correspondientes al grupo laterales-vibrantes ('ili iri iʎi irri' , 'ele ere eʎe erre', 'ala ara aʎa arra', 'olo oro oʎo orro', 'ulu uru uʎu urru').

Como se puede observar, la secuencia 'i e a o u' toma los colores rojo, verde, naranja, azul, morado. La primera conclusión importante, es que existe en este ejemplo una clara separabilidad entre vocales en el diagrama F1-F2, lo que facilita la identificación vocálica.

Aunque en los diagramas F1-F3, F2-F3 la separabilidad no es total, las superficies sólo tienen pequeños solapamientos, por lo tanto, esta información podría servir de ayuda en el caso de existir problemas en la identificación vocálica con el diagrama F1-F2.

En el diagrama F2-F3 se aprecia el descenso de frecuencias en ambos formantes que se produce en la secuencia estudiada ('i e a o u') tal y como cabría esperar.

Observando detenidamente las escalas, se puede establecer la conveniencia de introducir un nuevo gráfico que confronte bidimensionalmente los valores estudiados de la forma: eje x: (F2 menos F1), eje y: (F3 menos F1).

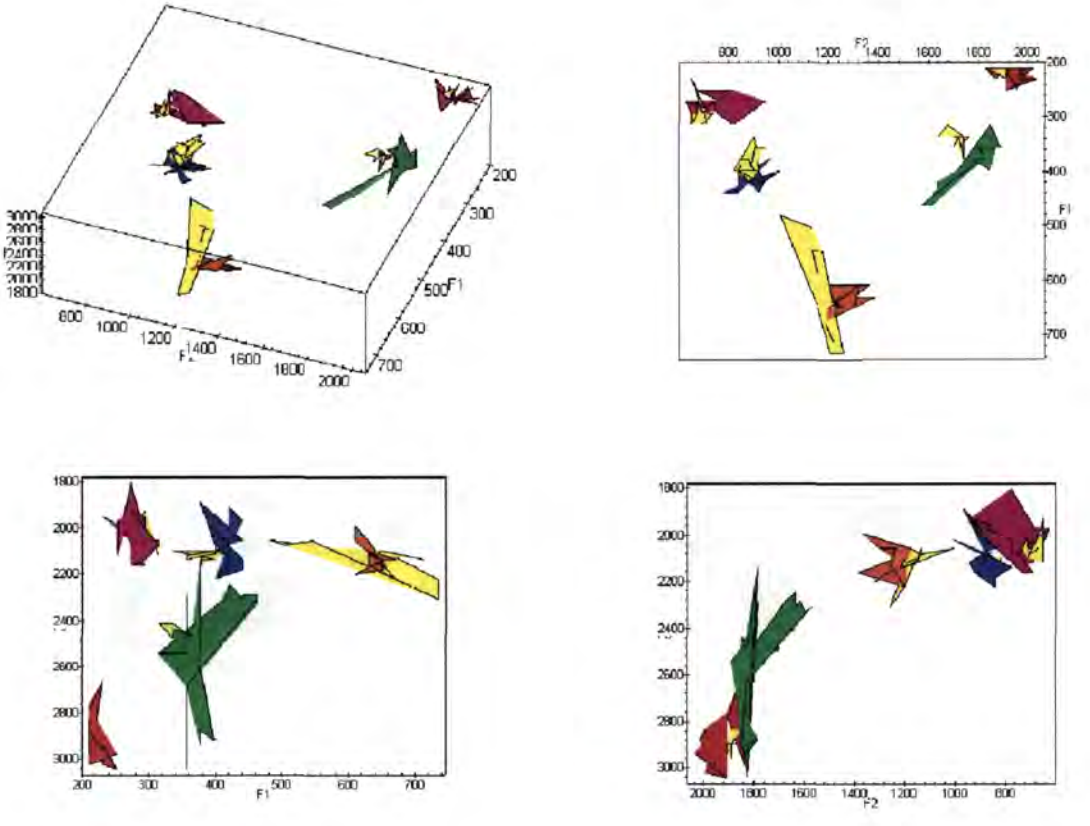


Figura 5.12
Posiciones de los tres primeros formantes de vocales junto a sonidos laterales-vibrantes confrontados entre sí.

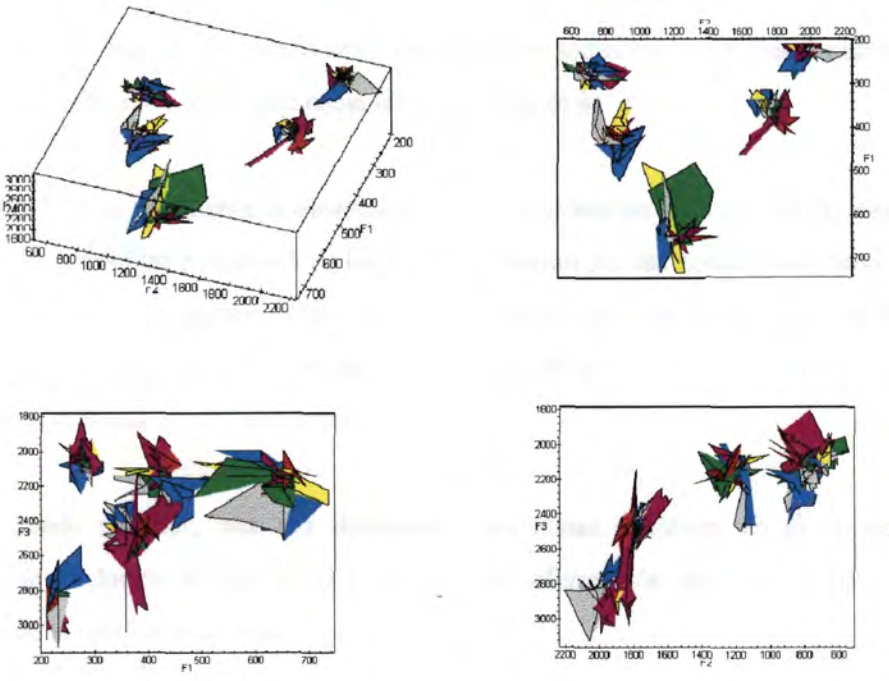


Figura 5.13

Posiciones de los tres primeros formantes de vocales frente a consonantes, cada color indica un grupo consonántico según el modo de articulación.

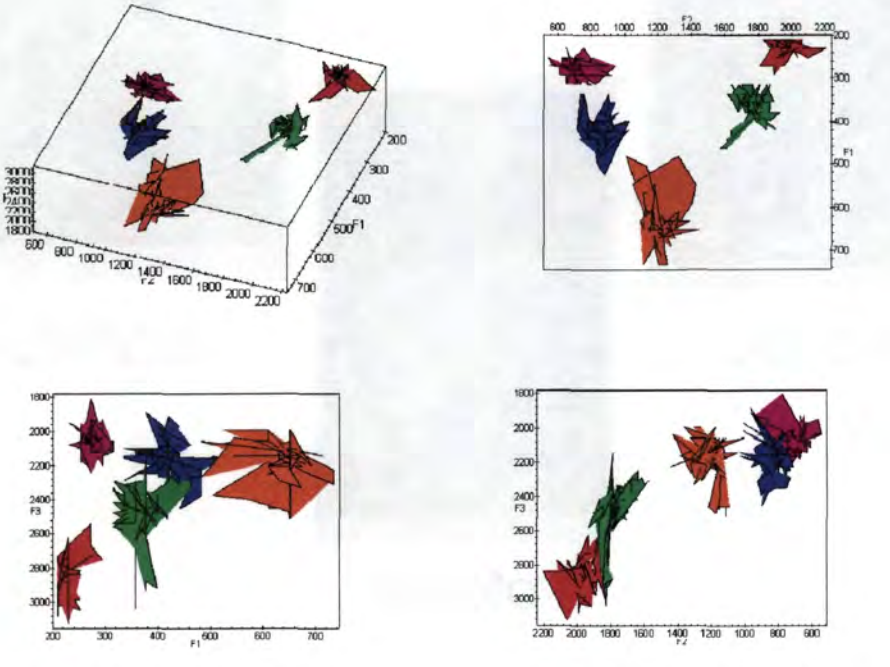


Figura 5.14

Posiciones de los tres primeros formantes de vocales frente a consonantes. Cada color indica una vocal (i:rojo, e: verde, a:naranja, o:azul, u:morado).

En la figura 5.15 se muestran los espectros de las palabras 'patético', 'picota' y 'ocupa'. La evolución de los formantes se aprecia muy claramente en todos los casos. Por ejemplo, en la 'a' de patético los formantes se dirigen hacia las posiciones de la 'e'.

En la figura 5.16 se representa la evolución de los formantes en cada una de las palabras. En 'ocupa' (color verde) las posiciones de los formantes pasan por las zonas típicamente ocupadas por la 'o', 'u' y 'a'. Lo mismo ocurre en 'picota' (color rojo) en donde se forma una figura aproximadamente triangular con vértices en las zonas de la 'i', 'o' y 'a'. 'Patético' (en color azul) también presenta el comportamiento esperado.

Como se puede apreciar, una vez delimitadas las zonas vocálicas en el espacio de los formantes, la evolución de los mismos nos da una información aproximada de las vocales contenidas en la oración analizada.

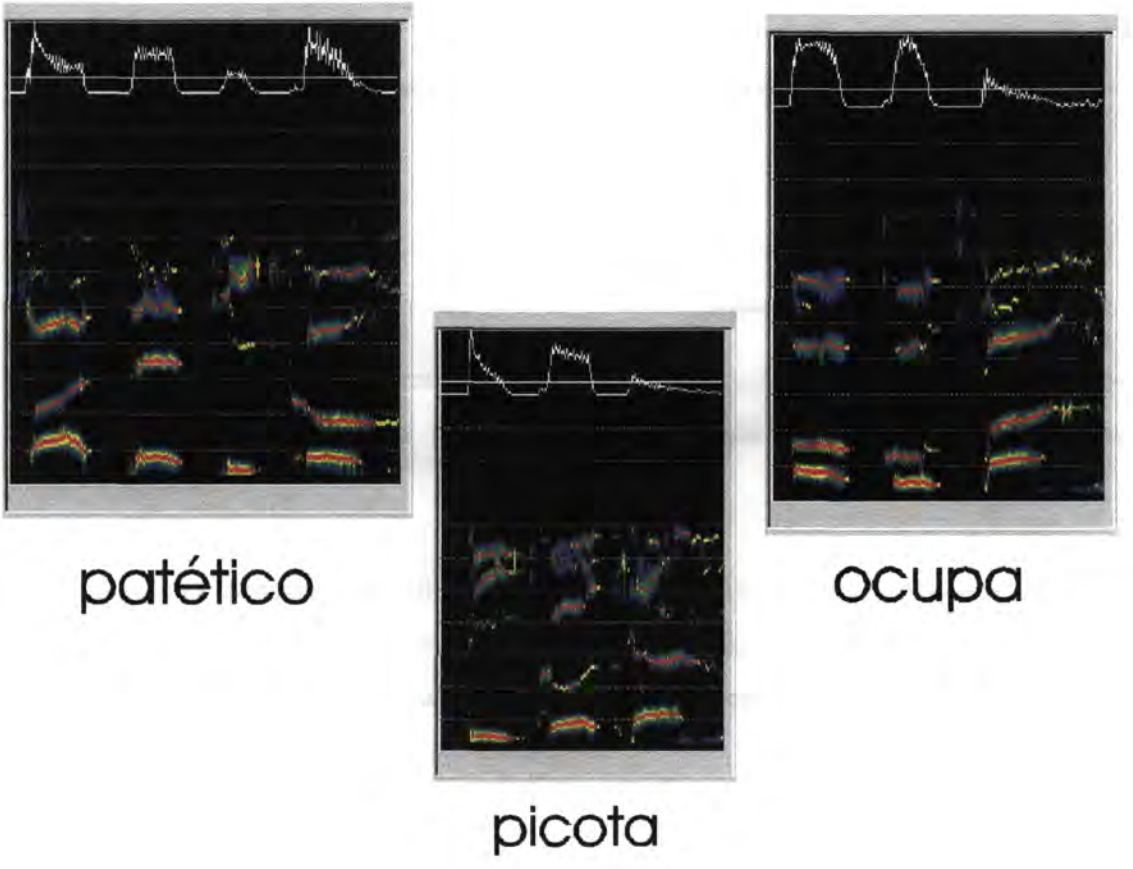


Figura 5.15
Espectros de voz de las palabras 'patético', 'ocupa', 'picota'.

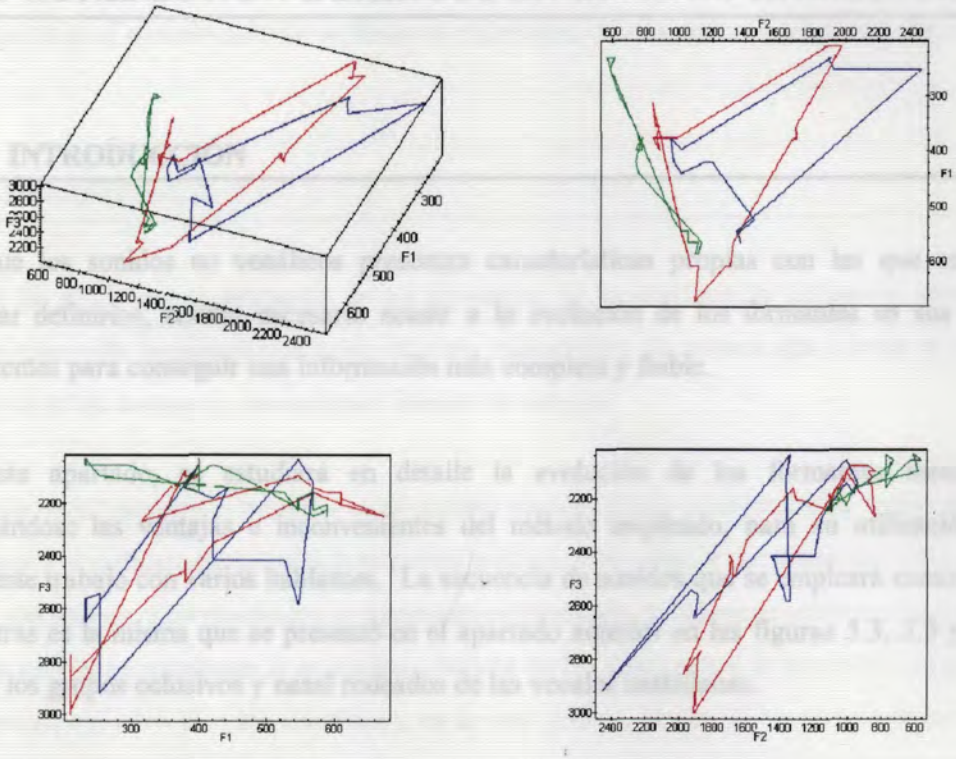


Figura 5.16

Evolución de los formantes en las palabras 'patético' (rojo), 'picota' (azul) y 'ocupa' (verde).

5.3.3 CONCLUSIONES

La representación visual de los formantes en planos de proyección (F1-F2, F1-F3, F2-F3), nos ofrece una visión complementaria a los espectros, que ayuda a clasificar e identificar los sonidos vocálicos.

Las vocales son fácilmente identificables gracias a la separación física existente entre las mismas en los planos de proyección utilizados. Los grupos oclusivos y nasal presentan características muy similares en cuanto a la posición y evolución de los formantes de las vocales adyacentes a las constantes que los integran.

La evolución de los formantes se puede representar en el tipo de gráficos propuesto, y será la base para los estudios del siguiente capítulo. También resulta necesario ampliar el conjunto de muestras analizadas con el fin de abarcar diversas variaciones del habla producidas por diferentes hablantes, con ello los resultados obtenidos serán más generales y fiables.

5.4 SONIDOS NO VOCÁLICOS EN UN SÓLO HABLANTE

5.4.1 INTRODUCCIÓN

Aunque los sonidos no vocálicos presentan características propias con las que se podría intentar definirlos, resulta necesario acudir a la evolución de los formantes en sus vocales adyacentes para conseguir una información más completa y fiable.

En este apartado, se estudiará en detalle la evolución de los formantes mencionada, sopesándose las ventajas e inconvenientes del método empleado, para su utilización en el siguiente trabajo con varios hablantes. La secuencia de sonidos que se empleará como base de muestras es la misma que se presentó en el apartado anterior en las figuras 5.3, 5.5 y 5.7, es decir, los grupos oclusivos y nasal rodeados de las vocales castellanas.

La razón por la que no se han empleado todos los sonidos posibles (faltan las consonantes laterales, fricativas y africada), es que el volumen y complejidad de los datos obtenidos hace más aconsejable restringirse a un subconjunto significativo de consonantes que nos permita obtener conclusiones y validar el método de análisis utilizado. El estudio completo se realizará al final de este capítulo en el apartado correspondiente a varios hablantes.

El análisis de los sonidos consonánticos rodeados de cada posible vocal, hace que el número de casos sea muy grande (90), aunque en este apartado sólo serán desarrollados la mitad de ellos. La ventaja de emplear este camino, es que una vez obtenidos los resultados, se dispone de la suficiente información como para poder destacar las características principales de la fonética acústica española, realizar generalizaciones, establecer reglas y destacar casos particulares.

Conviene aclarar que los sonidos que aquí se analizan están grabados por un hablante tipo (sin especiales cualidades o defectos articulatorios), se ha empleado un micrófono de calidad media, con un ruido de fondo bajo pero no despreciable, y la pronunciación no ha sido forzada. Todo ello, unido al hecho de que existe mucha variabilidad en los resultados, nos lleva a la conclusión de que las características espectrales que se obtendrán no deben tomarse como patrón, sino como referencia.

En el apartado en el que se realiza el análisis con varios hablantes, todos los inconvenientes expresados en el párrafo anterior se traducen en resultados más reales, generales y robustos.

5.4.2 DESARROLLO

Con el fin de obtener la evolución de los formantes, se ha ampliado la aplicación informática que se desarrolla paralelamente a los estudios presentados, de manera que calcule la posición espectral de los formantes en cualquier instante de tiempo.

De cada vocal anterior o posterior a un sonido consonántico, se seleccionan varios instantes espectrales significativos, se calcula la posición de sus formantes, y se realiza una interpolación lineal con splines. Los resultados se presentan comparando las cinco vocales en dos formatos, en una gráfica se representa la evolución absoluta de los formantes a lo largo del tiempo (eje x => tiempo, eje y => Hercios absolutos); en la otra gráfica, el eje 'y' representa el incremento positivo o negativo de la frecuencia (en Hercios) de un formante para cada vocal.

En la figura 5.17 se presenta la información correspondiente al sonido nasal bilabial 'm' del castellano, en cuanto a la evolución de los formantes de sus vocales adyacentes se refiere.

Las figuras 5.17 á 5.25 presentan una disposición interna idéntica, en la que las cuatro gráficas de la izquierda representan evoluciones del segundo formante y las cuatro de la derecha del formante tercero. Las 4 gráficas superiores muestran evoluciones referidas a valores absolutos en Hercios, y las cuatro inferiores se centran en las pendientes (diferencias de Hercios) de los formantes a lo largo del tiempo (medido en milisegundos). Por último, las gráficas situadas en filas impares (primera y tercera) representan las vocales que preceden al sonido consonántico, mientras que las situadas en las filas segunda y cuarta muestran los datos referentes a las vocales que siguen al sonido consonántico.

La gráfica superior izquierda de la figura 5.17, muestra la evolución en términos absolutos del segundo formante de las vocales precediendo al sonido 'm'. En éste y todos los demás casos, la 'i' se representa con el color rojo, la 'e' con el verde, 'a' azul claro, 'o' azul oscuro, 'u' amarillo. Todas las vocales se encuentran en el margen de frecuencias que les corresponde, lo que se puede comprobar comparando las posiciones espectrales con las presentadas en la figura 5.14 del apartado anterior.

La evolución de los formantes se aprecia mucho mejor en la tercera gráfica (empezando desde arriba) de la izquierda, correspondiente al progreso en términos relativos del segundo formante en vocales precediendo al sonido [m]. Se puede determinar con claridad como la tendencia general es descendente, debido al locus bajo que imponen los sonidos bilabiales.

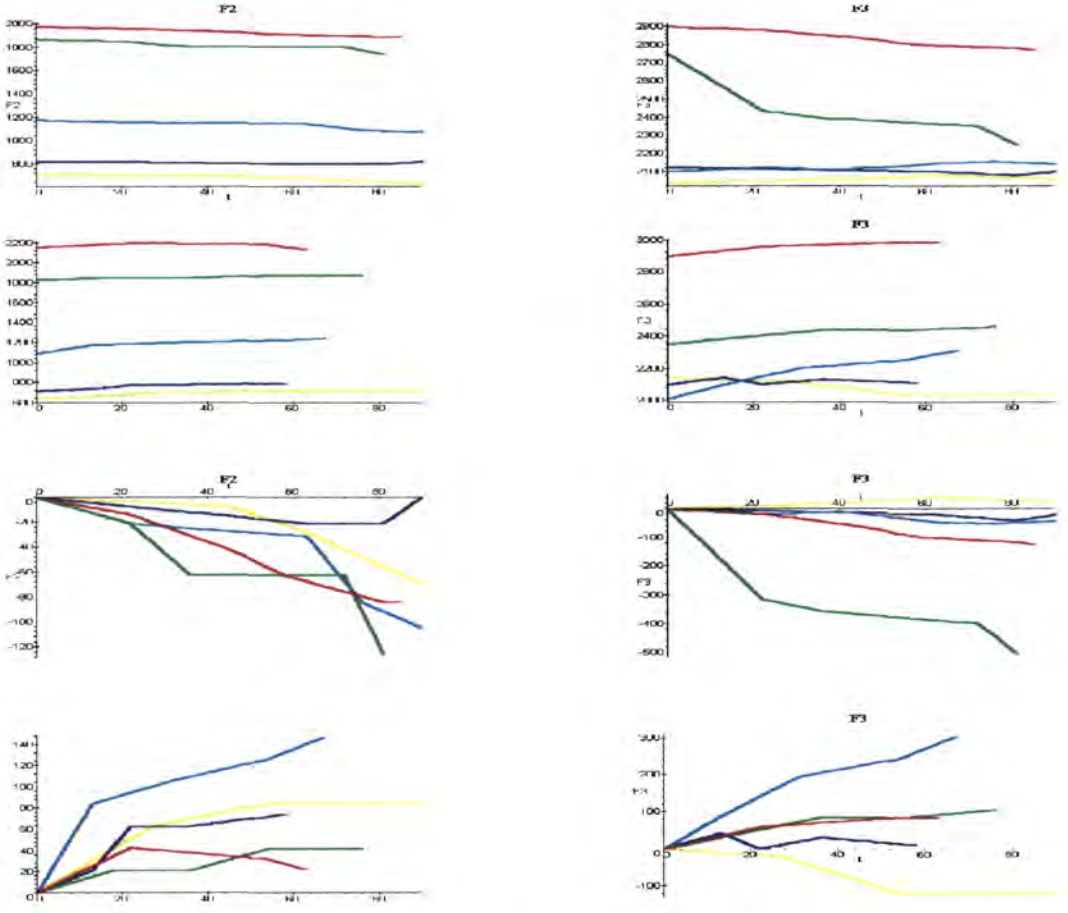


Figura 5.17
Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido nasal [m].

F2: Gráficas de la izquierda. F3: Gráficas de la derecha.
Evoluciones absolutas: Gráficas superiores. Relativas: Gráficas inferiores
Funciones: 'i' : rojo, 'e': verde, 'a': azul claro, 'o': azul, 'u': amarillo
Filas 1 y 3: vocales anteriores. Filas 2 y 4: vocales posteriores

La segunda y cuarta gráfica de la izquierda son las que representan las posiciones absolutas y evoluciones del segundo formante en las vocales que se hallan tras el sonido [m]. En este caso, también se aprecia con mucha claridad la subida prevista de los formantes partiendo de un locus bajo.

Las gráficas de la columna de la derecha representan exactamente lo mismo que las de la columna de la izquierda, pero referidas al tercer formante, que en general es más inestable y variable que el segundo. También aquí se aprecian las tendencias generales esperadas.

La figura 5.18 presenta en la vocal anterior al sonido [n] una clara elevación en las vocales 'o' y 'u', mientras que en estas mismas vocales existe una bajada cuando se encuentran después del sonido consonántico. Esto indica la situación de un locus de nivel medio, más alto en frecuencia que el del sonido bilabial.

La tendencia que exhibe el segundo formante se aprecia mejor en las gráficas inferiores, donde se confirma la idea expresada en el párrafo anterior, salvo en las vocales con frecuencias más altas (i,e), especialmente la 'e', que reafirma la idea de un locus de nivel medio propio de los sonidos dentales-alveolares.

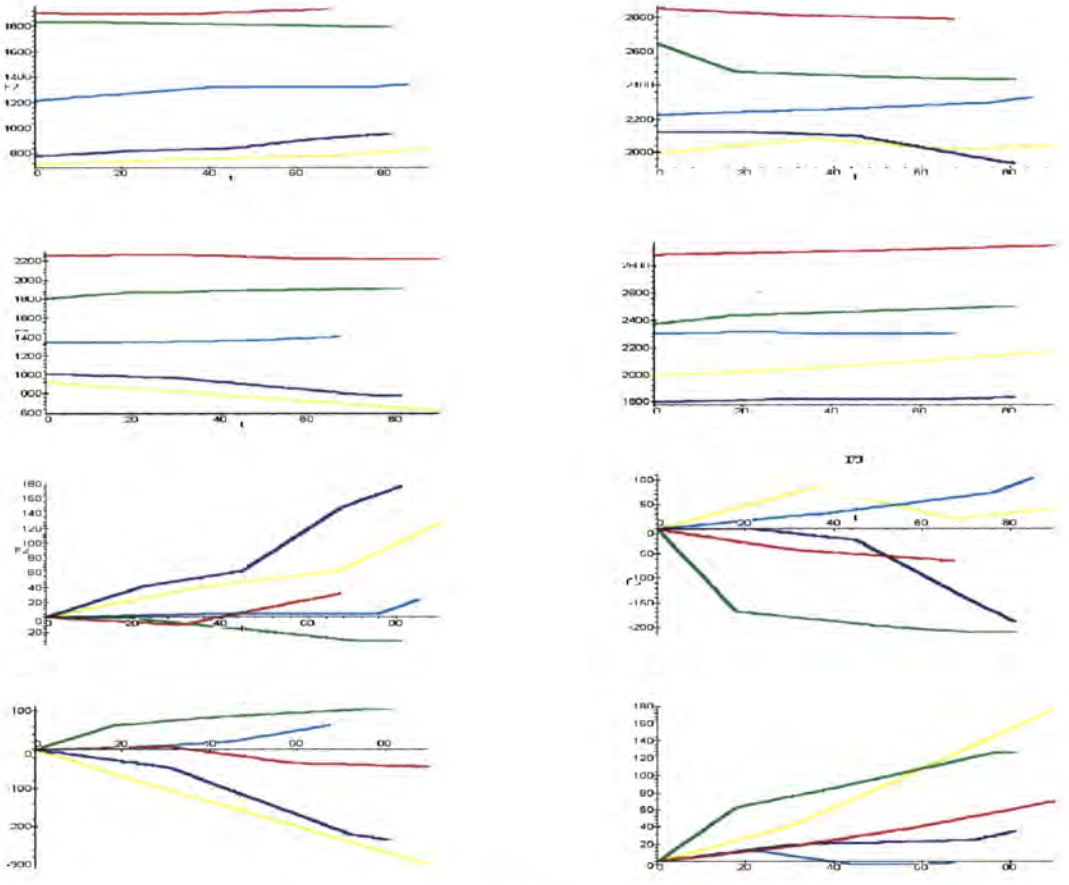


Figura 5.18

Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido nasal [n].

El tercer formante, situado más alto en frecuencias, se ve obligado a evolucionar hacia el locus de forma más brusca que el segundo. Existe una tendencia hacia la bajada en las vocales anteriores a [ɳ] y subidas claras en las vocales posteriores.

En la figura 5.19 correspondiente al sonido velar [ŋ] , la evolución de los formantes es muy pronunciada, siguiéndose las pautas esperadas hacia un locus posicionado en altas frecuencias, por lo que, como se puede apreciar, las pendientes de las curvas son inversas al caso bilabial. En las vocales anteriores existen subidas y en las posteriores bajadas.

El comportamiento teórico del tercer formante no coincide totalmente con el esperado, pero como se verá en el apartado dedicado a varios hablantes, existen diversas formas de manifestación de las características espectrales de éste y otros sonidos.

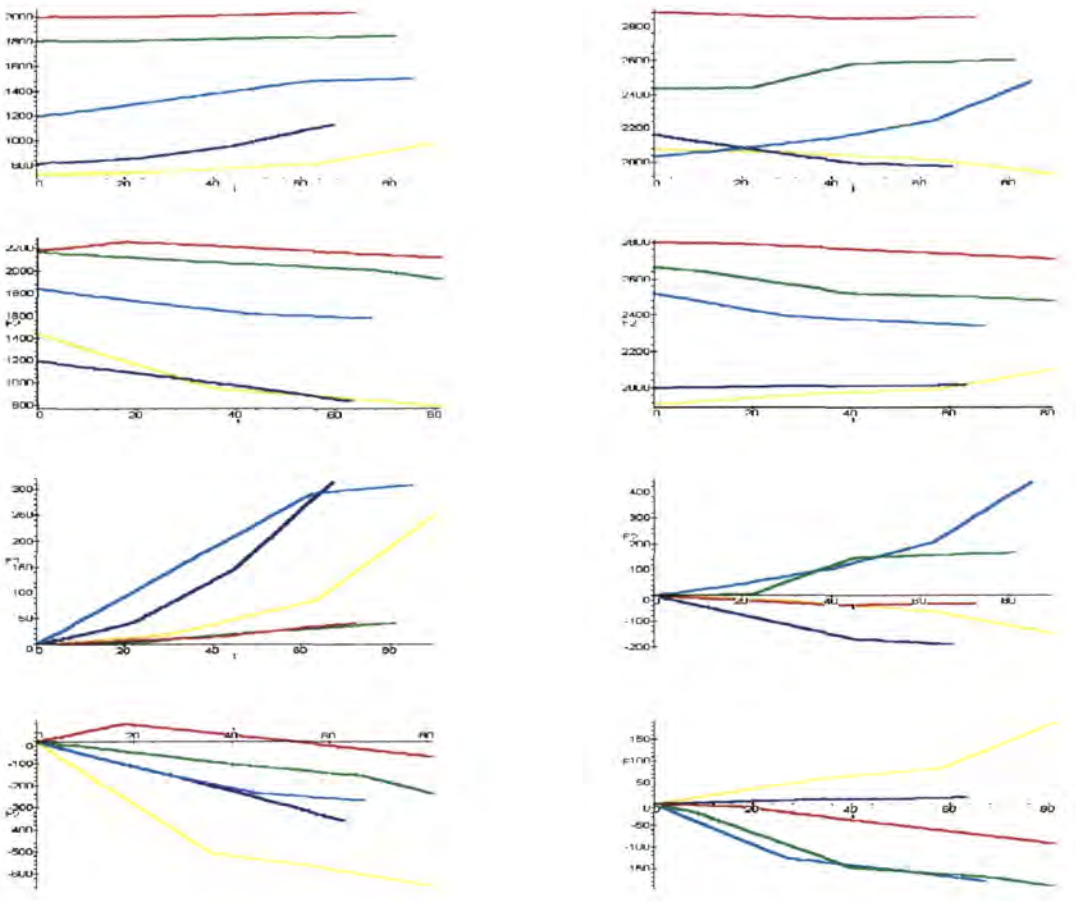


Figura 5.19
Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido nasal [ɳ].

Las próximas tres figuras cubren el grupo oclusivo sordo. En el primer caso (figura 5.20), se analiza el sonido bilabial [p]. Como cabe esperar, las evoluciones de los formantes se asemejan al también bilabial [m] anteriormente analizado.

Los formantes F2 y F3 descienden en este caso hacia un locus bajo en frecuencia. En el ejemplo, las vocales posteriores a la consonante muestran especialmente claro este descenso en ambos formantes (gráficas inferiores).

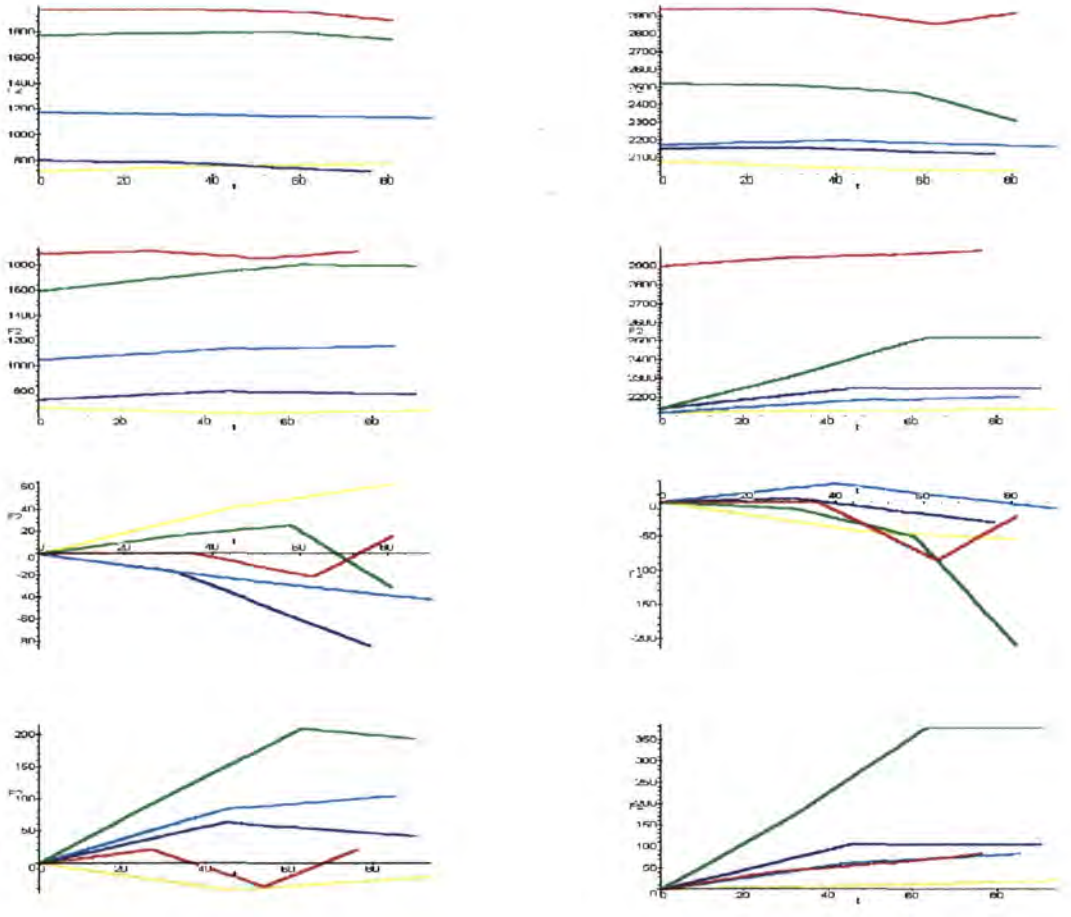


Figura 5.20

Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido oclusivo sordo [p].

En la figura 5.21 correspondiente al sonido [t], al igual que en [n], existe un locus medio, y esto se manifiesta en las bajadas de los formantes de las vocales 'i', 'e', 'a'. Sin embargo, en las vocales 'o' y 'u' donde el segundo y tercer formante son bajos, éstos tienden a subir hacia el locus.

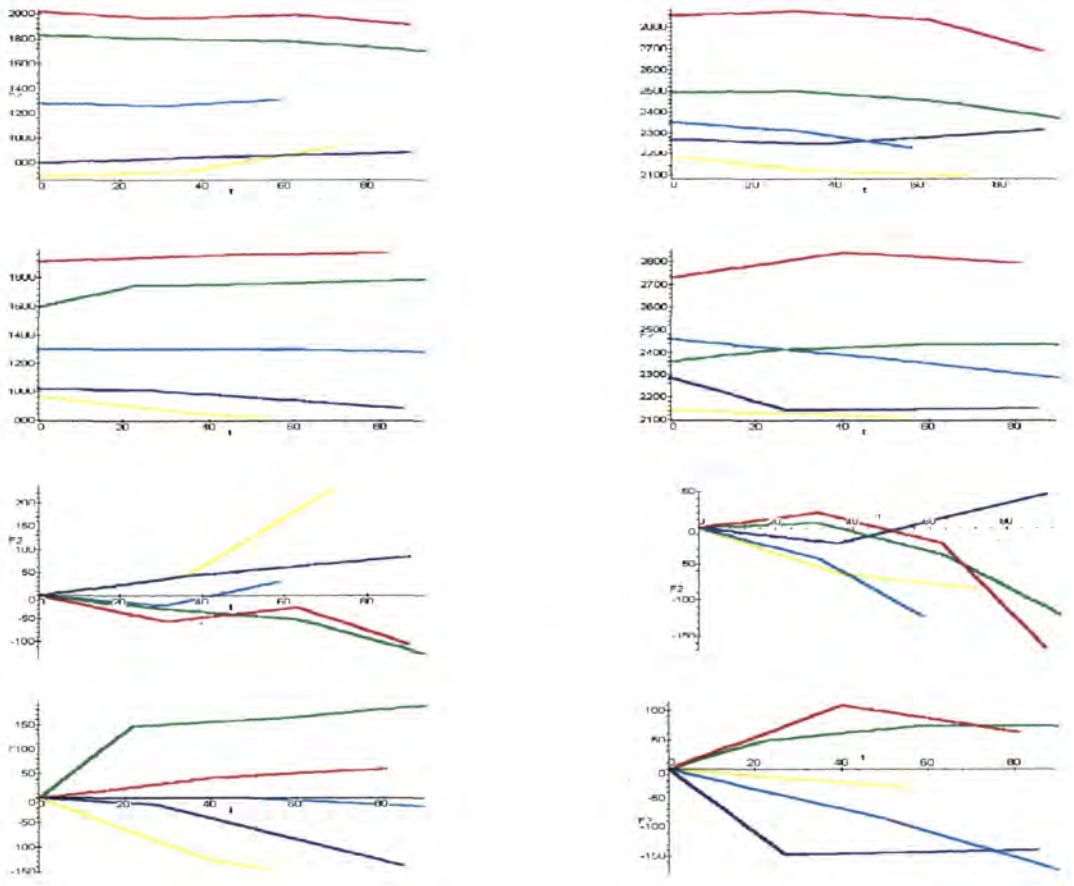
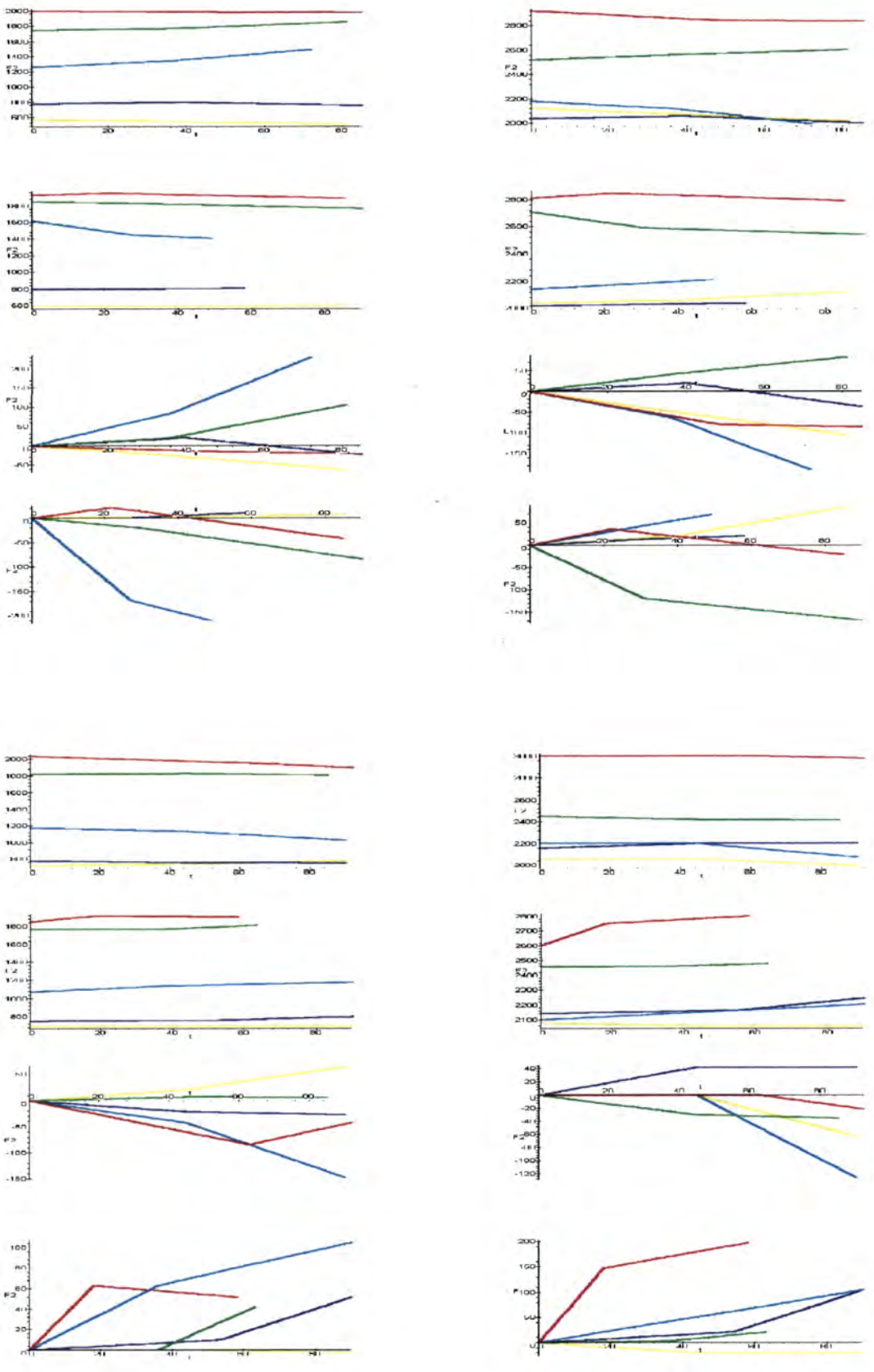


Figura 5.21
Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido oclusivo sordo [t].

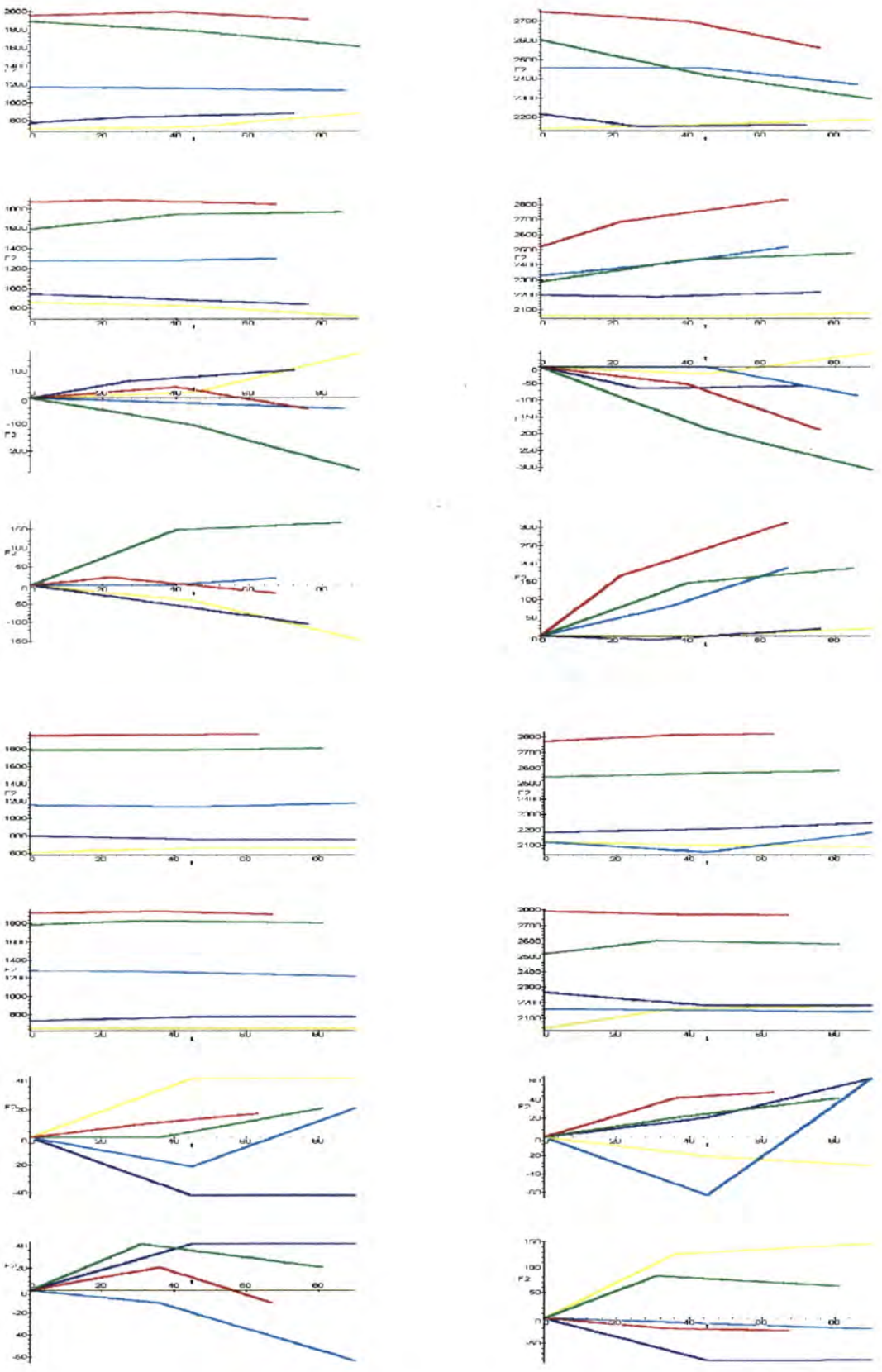
La figura 5.22 cierra el grupo oclusivo sordo, el sonido representado es el equivalente a la consonante 'k'. Las funciones presentan el comportamiento típico de los sonidos velares, en donde F2 se eleva hacia el locus, mientras que F3 baja, tendiendo ambos a juntarse en los extremos del sonido consonántico. Esta característica se aprecia en las gráficas por la simetría que existe respecto al eje x en la pendiente de los formantes de las vocales anterior y posterior al sonido consonántico.

Las figuras 5.23, 5.24 y 5.25 abarcan el grupo fricativo sonoro [β], [d.], [γ], sus características son similares a los sonidos anteriores atendiendo al punto de articulación, de manera que, por ejemplo, [β] es parecido en la evolución de sus formantes a [p] y [m], y análogamente con los demás sonidos.



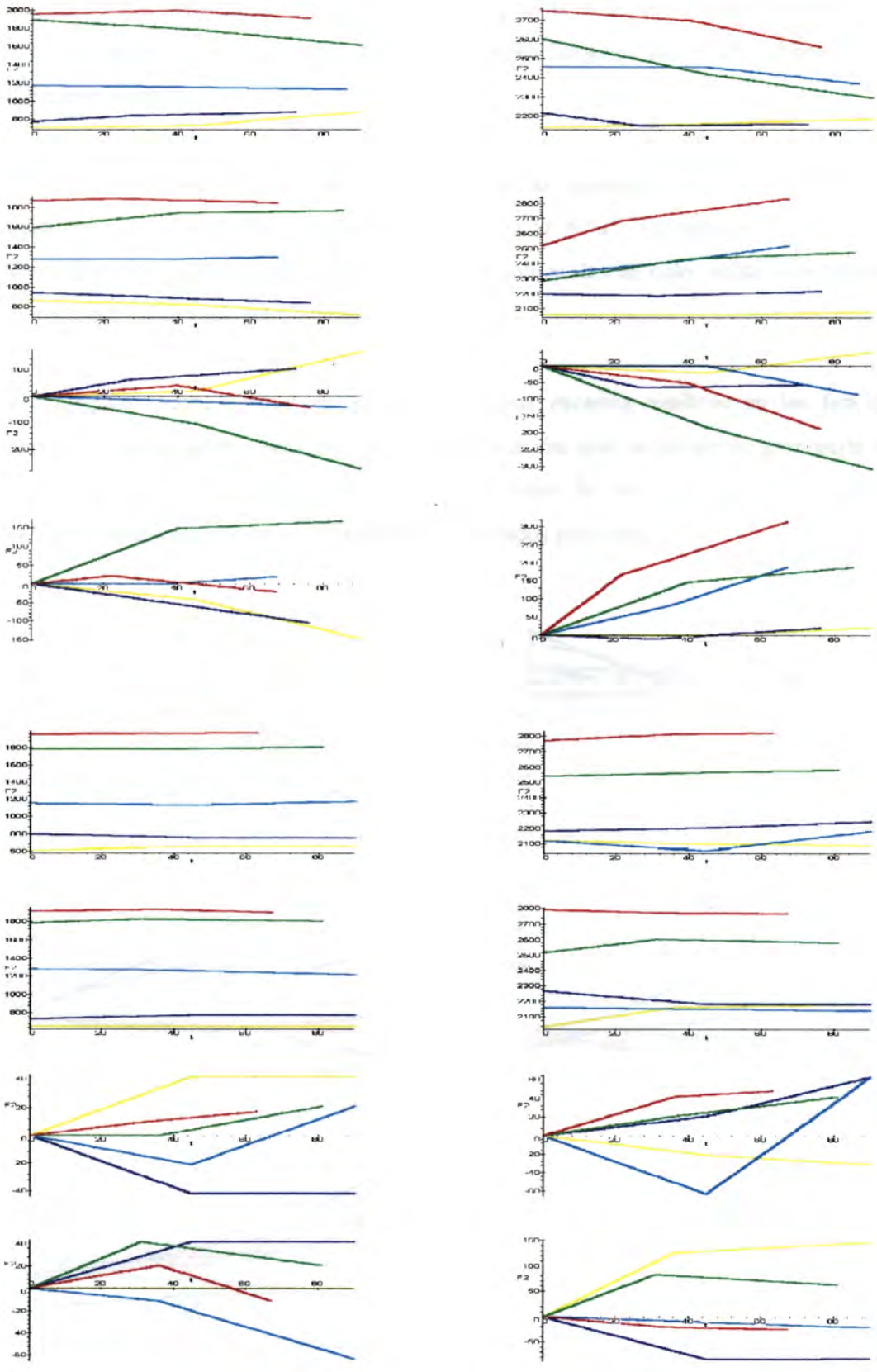
Figuras 5.22 y 5.23

Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes al sonido oclusivo sordo [k] y fricativo sonoro [β].



Figuras 5.24 y 5.25

Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes a los sonidos fricativos sonoros [d.] y [γ].



Figuras 5.24 y 5.25

Evoluciones absolutas y relativas de F2 y F3 en las vocales adyacentes a los sonidos fricativos sonoros [d.] y [γ].

La figura 5.26 reúne todas las funciones que representan la evolución del segundo formante de las vocales que preceden a cada una de las consonantes del grupo nasal, del oclusivo sordo y del fricativo sonoro.

Cada gráfica representa a una vocal ('i': gráfica superior izquierda, 'e': superior derecha, 'a' izquierda en la siguiente fila, y así sucesivamente con 'o' y 'u'). Las funciones de color rojo indican la evolución de F2 ante sonidos oclusivos sordos, las de color verde ante fricativos sonoros y por fin en azul, ante nasales.

La conclusión que se obtiene, es que no existe una excesiva similitud en las funciones comparándolas por grupos según el modo de articulación, esto es debido en gran parte a la variabilidad de los resultados en distintas realizaciones de voz. Sin embargo, se pueden establecer pautas generales que se detallarán en apartados posteriores.

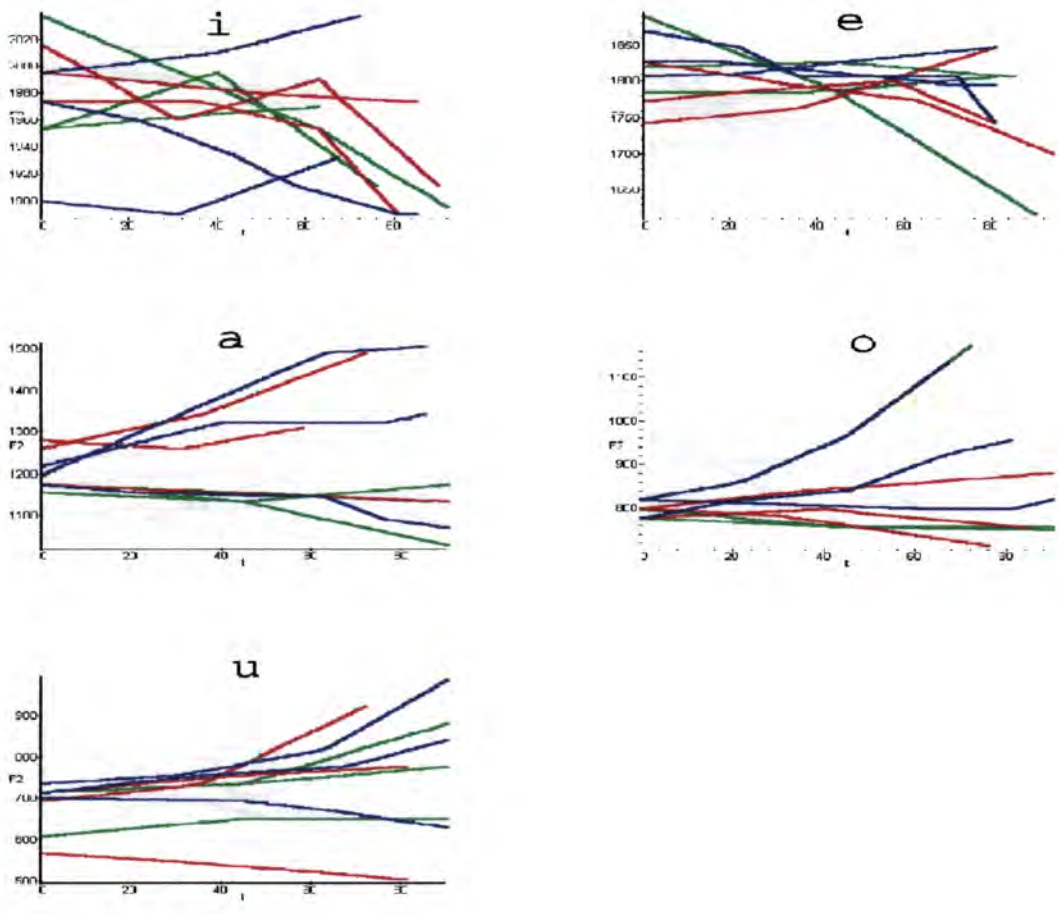


Figura 5.26

Evolución de F2 en las vocales situadas ante consonantes de tipo nasal (azul), oclusivo sordo (rojo) y fricativo sonoro (verde).

Otra conclusión importante se basa en la constatación de que el rango de frecuencias que puede adoptar cada vocal varía fundamentalmente en los extremos de las mismas; esto es debido a la evolución de los formantes hacia los locus de las consonantes, que a veces se encuentran bastante alejados de las frecuencias típicas de algunas vocales, como por ejemplo una 'i' ante consonante bilabial, o una 'o' ante velar.

La densidad que existe en cada una de las gráficas, nos da una idea de cual es la media y la varianza en Hercios del segundo formante para cada una de las vocales.

La figura 5.27 tiene el mismo significado que la figura 5.26, pero en este caso todas las funciones representan al tercer formante. Los razonamientos y observaciones expresados para F2 son extensibles a F3, aunque en este caso, existe una mayor dispersión (varianza) en el rango de valores frecuenciales que adoptan las funciones dibujadas.

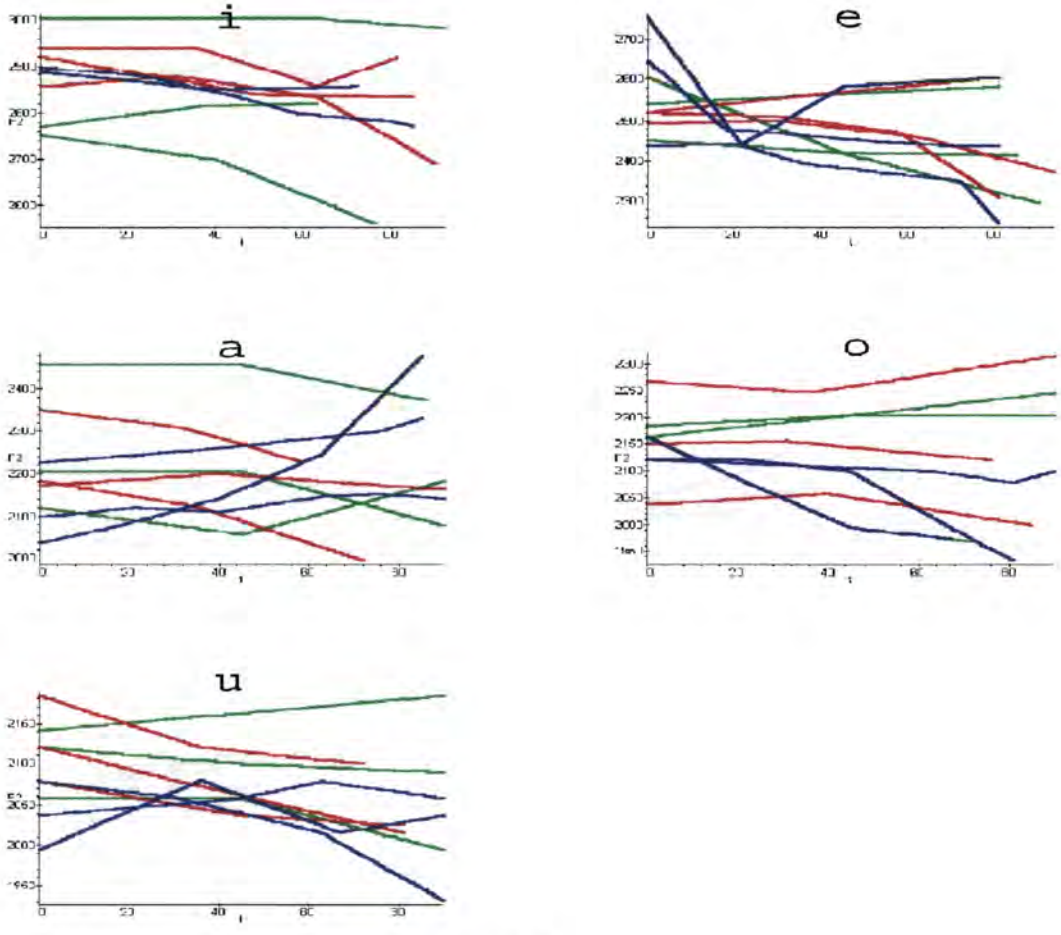


Figura 5.27

Evolución de F3 en las vocales situadas ante consonantes de tipo nasal (azul), oclusivo sordo (rojo) y fricativo sonoro (verde).

La figura 5.28, al igual que la figura 5.26, muestra la evolución de F2 en vocales seguidas de sonidos consonánticos nasales y oclusivos. En este caso, las funciones representadas se agrupan atendiendo al punto de articulación de la consonante. El color rojo hace referencia a vocal ante bilabial, el verde ante dental-interdental, y el azul ante velar.

De nuevo, los resultados no son 'ejemplares' en el sentido de encontrarnos evoluciones perfectas hacia locus posicionados en las alturas esperadas, o comportamientos idénticos en distintas vocales con el mismo punto de articulación, sin embargo, sí se puede hablar de tendencias, como por ejemplo en la 'i' (gráfica superior izquierda), donde el grupo bilabial fuerza evoluciones descendentes y el grupo velar crecientes o mantenidas.

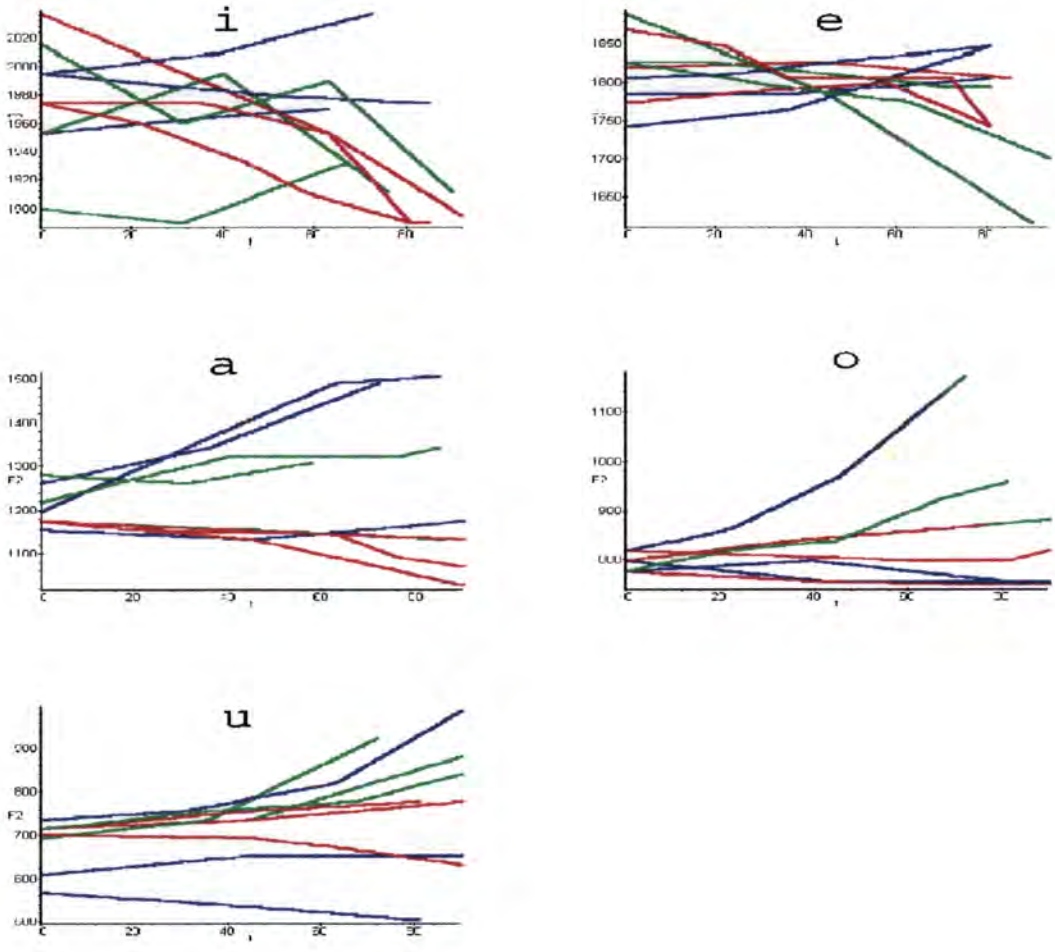


Figura 5.28

Evolución de F2 en las vocales situadas ante consonantes de tipo bilabial (rojo) , dental-interdental (verde) y palatal-velar (azul).

En la figura 5.29 las funciones representan a F3. Al igual que en el caso del segundo formante, se aprecia un mayor agrupamiento (similitud) en las transiciones forzadas por la presencia de consonantes bilabiales.

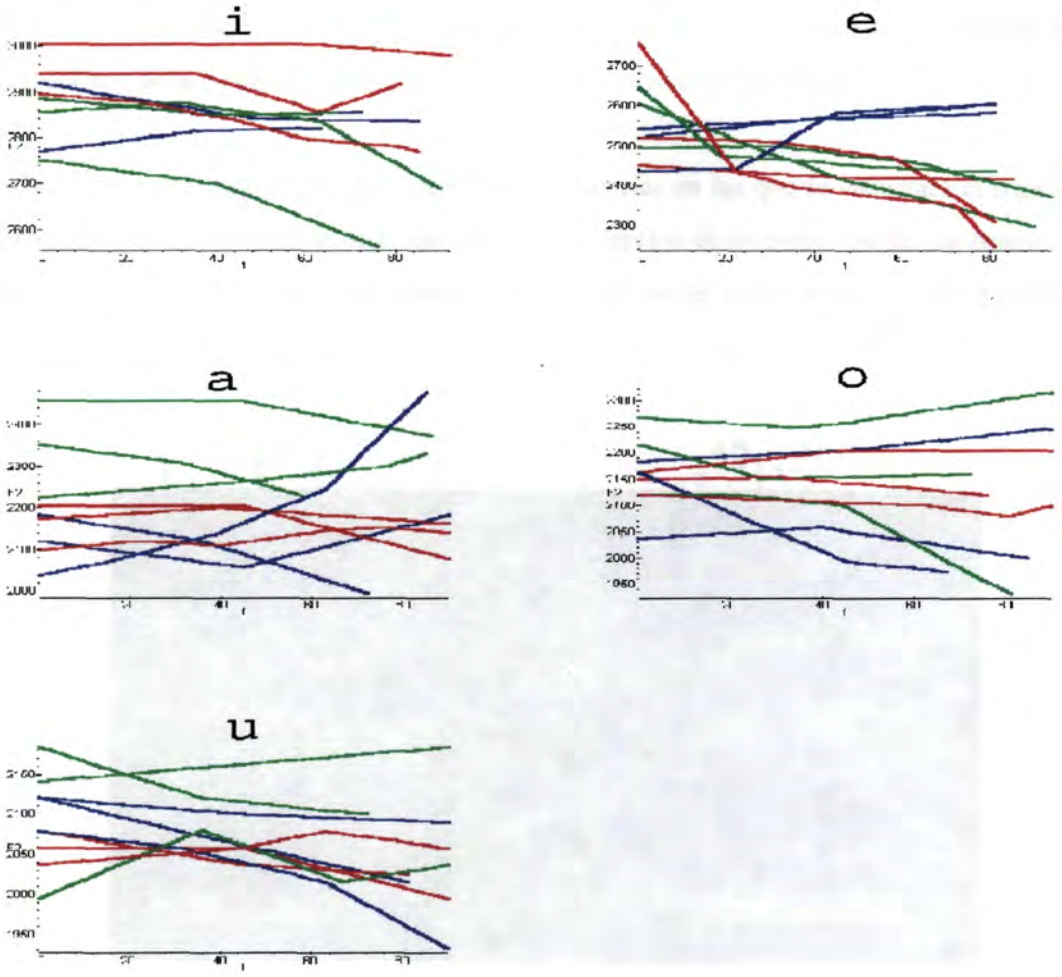


Figura 5.29
Evolución de F3 en las vocales situadas ante consonantes de tipo bilabial (rojo), dental-interdental (verde) y palatal-velar (azul).

El tipo de análisis seguido en este apartado resulta muy laborioso, y las funciones obtenidas por interpolación no son fáciles de comparar entre sí para generalizar resultados, por ello se ha considerado más adecuado seguir una metodología en la que en primer lugar se identifiquen visualmente las variaciones espectrales que habitualmente se presentan para cada sonido, y tras ello, se especifiquen las posiciones y evoluciones típicas de los formantes para cada una de las situaciones más representativas que han sido identificadas.

A modo de ejemplo, en las figuras 5.30 y 5.31 se presentan varias gráficas de los sonidos 'epe' y 'eke'. En la figura 5.30, correspondiente al sonido 'epe', se aprecia en los tres casos representados la misma evolución, que se corresponde con el comportamiento esperado; sin embargo, como es lógico, la duración de las vocales y la brusquedad de las evoluciones varía de una gráfica a otra, dependiendo de la velocidad en el habla que hubo en el momento de la grabación y de la posición de los órganos articulatorios en ese momento .

Sobre el espectro superior, se han escrito las frecuencias en las que se encuentra el segundo y tercer formante, así como la distancia medida en Hercios de la evolución de sus transiciones hacia el locus bilabial. También aparecen las duraciones (en milisegundos) más significativas de cada porción del espectro.

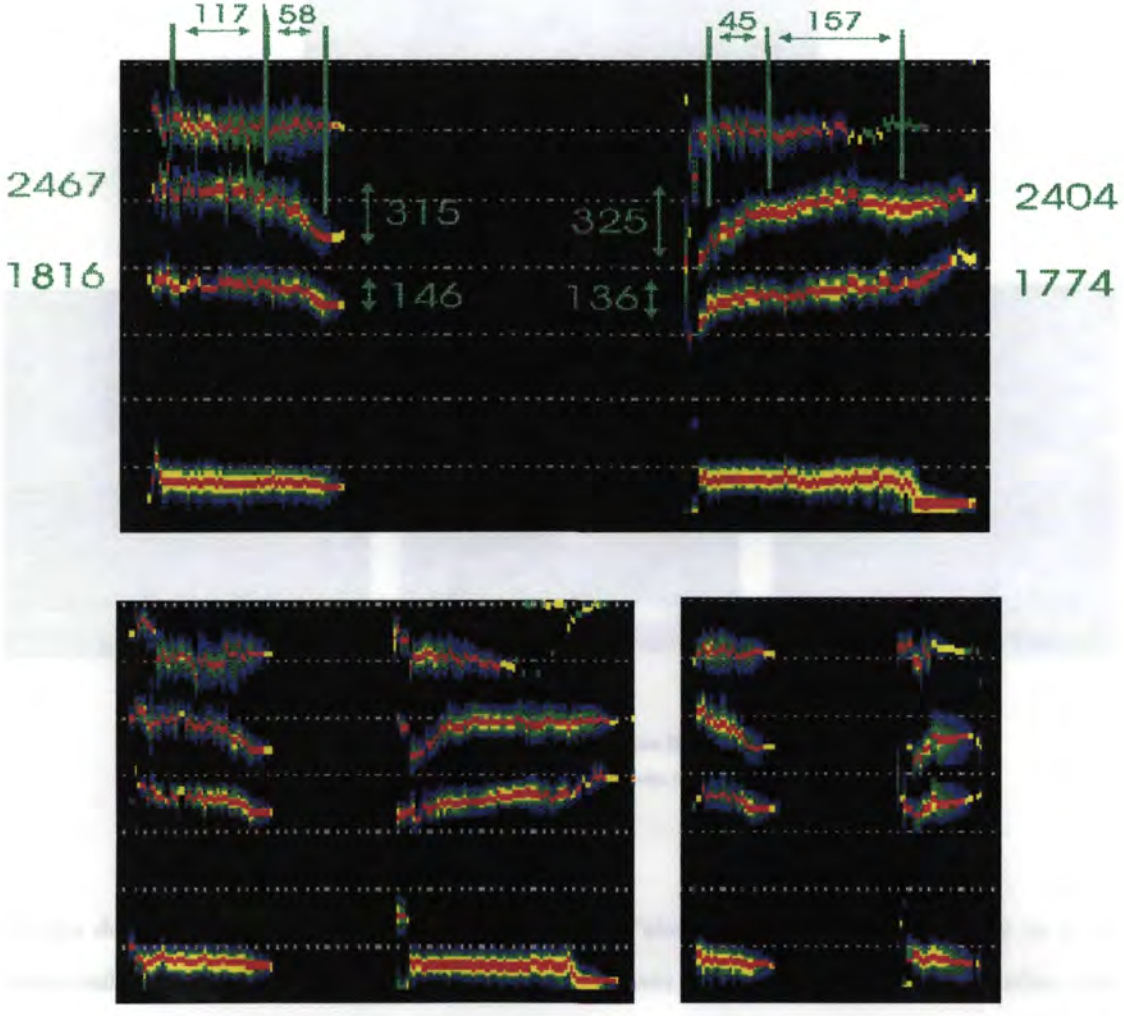


Figura 5.30

Espectros típicos del sonido 'epe' con valores (en Hercios) de la posición y evolución de los formantes segundo y tercero.

La figura 5.31 muestra algunas variaciones espectrales típicas con las que aparece el sonido 'eke'. Una situación habitual es que el locus se encuentre entre la altura del segundo y el tercer formante. También está representada en las gráficas inferiores de la derecha, la situación en la que el locus se encuentra más alto que el tercer formante, con lo que existen subidas de F2 y F3 en las vocales anterior y posterior al sonido velar.

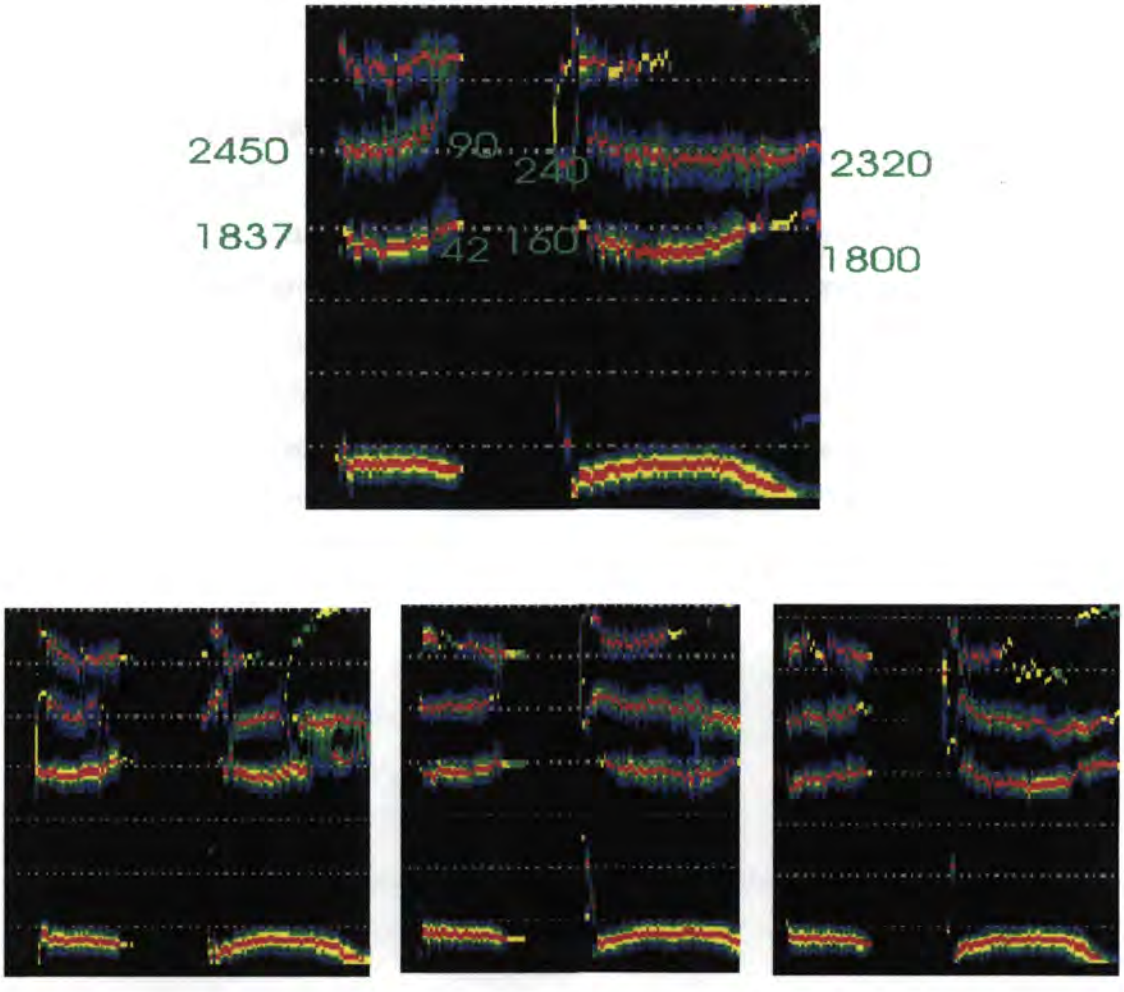


Figura 5.31
Espectros típicos del sonido 'eke' con valores (en Hercios) de la posición y evolución de los formantes segundo y tercero.

El tipo de estudio realizado con los sonidos 'epe' y 'eke' para un sólo hablante, no va a ser continuado, con el fin de evitar duplicaciones extensas e innecesarias entre apartados. Las características de las figuras 5.30 y 5.31, junto con sus explicaciones, nos sirven de referencia para habituarnos a la manera en que se presentarán los resultados en el apartado dedicado al análisis de los sonidos no vocálicos en varios hablantes.

5.4.3 CONCLUSIONES

Para poder caracterizar los espectros de voz, es necesario comprender la evolución de los formantes en las vocales adyacentes a los sonidos consonánticos, sin embargo, éstas evoluciones, aunque siguen unas reglas y pautas generales, presentan mucha variación entre distintas realizaciones del habla.

Los márgenes de frecuencias de los formantes que presenta cada vocal no varían significativamente según cual sea el contexto consonántico, salvo en las porciones en las que existe evolución hacia los distintos locus posibles.

Para poder realizar análisis representativos del habla, es necesario obtener, confrontar y comparar un gran número de grabaciones de voz de diferentes hablantes pronunciando sonidos básicos rodeados de diferentes contextos. Esta exigencia, junto al hecho de la fuerte variabilidad que presentan los resultados, nos lleva a la elección de una metodología de análisis de espectros basada en generalizaciones de características de 'grano grueso', cuyos parámetros sean únicamente posiciones y variaciones frecuenciales básicas de los formantes.

5.5 SONIDOS VOCÁLICOS EN VARIOS HABLANTES

5.5.1. INTRODUCCIÓN

Determinar el posible rango de frecuencias que puede adoptar cada formante en las vocales castellanas, ayudaría a la comprensión de los espectros de voz y al reconocimiento de sonidos aislados basado en la localización de sus formantes constitutivos. Este objetivo requiere la utilización de una base de datos de muestras lo suficientemente grande y variada como para qué garantice la generalidad de los resultados.

El estudio que aquí se presenta, pretende ser una referencia de como clasificar, obtener y procesar los datos de entrada (muestras de voz), con el fin de conseguir unos resultados claros y representativos. El número de hablantes empleados ha sido cuatro (dos hombres y dos mujeres), lo que hace imposible considerar los resultados como definitivos a pesar de las más de 3000 muestras de formantes utilizadas. Con las herramientas informáticas desarrolladas se podría abordar con facilidad un estudio más amplio que aunque resultaría tedioso, sería sencillo de llevar a cabo.

5.5.2 METODOLOGÍA EMPLEADA

Las bases de actuación se establecieron en el apartado correspondiente a un sólo hablante, sin embargo, debido a la poca incidencia habida clasificando las vocales según sus consonantes adyacentes agrupadas por el modo de articulación, en este caso se ha decidido realizar un trabajo similar, pero cambiando los grupos consonánticos de tal manera que se clasifiquen por el punto de articulación.

Las muestras han sido tomadas para cada uno de los hablantes completando los siguientes grupos:

1.- Vocales adyacentes a sonidos bilabiales

2.- Vocales adyacentes a sonidos dentales e interdentales

3.- Vocales adyacentes a sonidos velares

4.- Vocales aisladas

Las consonantes seleccionadas han sido las nasales, oclusivas sordas y fricativas sonoras.

Las etapas fundamentales para la realización de este trabajo son:

1.- Grabación de las muestras, agrupadas atendiendo a:

a) Hablante
b) Grupo (bilabial, dental/interdental, palatal/velar, aisladas)
c) vocal (a,e,i,o,u)

2.- Cálculo de los parámetros de predicción lineal de cada grabación.

3.- Obtención de los espectros de voz.

4.- Determinación de la posición de los formantes mediante los algoritmos de tratamiento de señal desarrollados en el capítulo anterior.

5.- Selección de los instantes de tiempo más significativos en cada grabación.

6.- Inserción ordenada en una base de datos de los valores frecuenciales de los tres primeros formantes para todos los instantes de tiempo seleccionados en la etapa anterior.

7.- Realización de selecciones sobre la base de datos, atendiendo a diferentes combinaciones de los parámetros establecidos en la primera etapa.

8.- Creación de mapas tridimensionales que sitúen a los sonidos en el espacio F1, F2, F3.

9.- Proyección de cada mapa tridimensional sobre los tres planos posibles (F1-F2, F1-F3, F2-F3).

10.- Agrupación y visualización de resultados.

La figura 5.32 presenta tres ejemplos de espectros obtenidos tras la consecución del paso 4. En este caso, se muestra en la gráfica superior izquierda los formantes realzados correspondientes a los sonidos 'opo', 'oβo', 'omo', todos ellos bilabiales. En la gráfica superior derecha se representa la secuencia 'oto', 'od.o', 'ono' y en la inferior las velares 'oko', 'oγo', 'oŋo'.

Como se puede apreciar, la evolución de los formantes se presenta muy clara, y será tema de estudio en la próxima sección de este capítulo, centrándonos ahora únicamente en las posiciones frecuenciales, sin incidir en las variaciones que existen a lo largo del tiempo.

De cada uno de los espectros con formantes realzados conseguidos en el paso 4, se seleccionan en la etapa 5 unos 15 instantes de tiempo significativos, lo que nos lleva a la obtención de unos 3600 valores espectrales de formantes (15 x 5 vocales x 4 grupos x 4 hablantes x 3 formantes).

El resultado final del trabajo se representa mediante visualizaciones de los sonidos sobre planos obtenidos tras realizar proyecciones que parten del espacio de los formantes F1, F2, F3 (iguales a las utilizadas en el apartado dedicado a un sólo hablante).

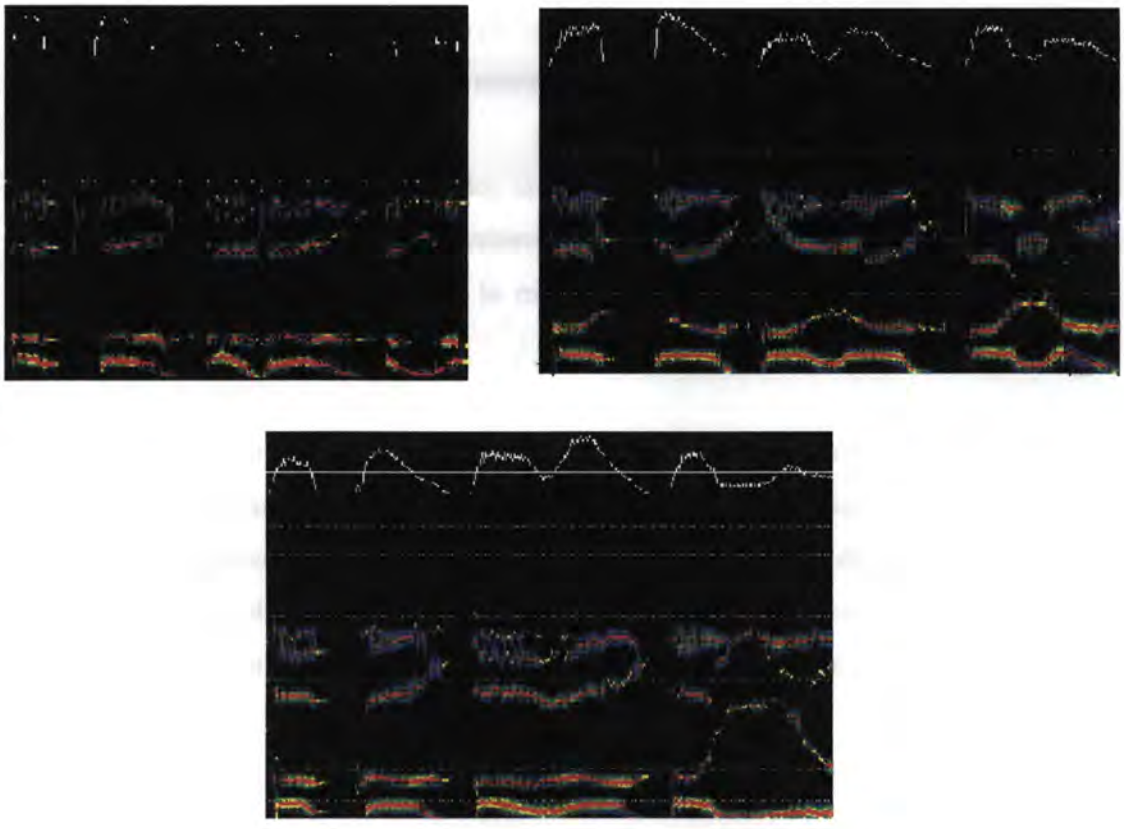


Figura 5.32
Ejemplo de espectros sobre los que se obtienen los valores de los formantes.

5.5.3 RESULTADOS OBTENIDOS EN CADA UNO DE LOS HABLANTES

En este apartado se presentan parte de los resultados que se pueden extraer para cada uno de los 4 hablantes tomados como referencia. En el primero de ellos, correspondiente al sexo masculino, se detallarán en gráficas separadas las posiciones de las vocales castellanas para cada uno de los grupos articulatorios escogidos. En el resto de los hablantes se presentan únicamente los resultados finales, con el objetivo de simplificar la exposición del estudio.

En la figura 5.33, aparecen las posiciones frecuenciales de los formantes correspondientes a las vocales aisladas del primer hablante. Estas posiciones nos servirán de referencia para poder comparar las desviaciones que se dan en los tres grupos de estudio restantes.

La característica más significativa es la excelente separabilidad que existe para todas las vocales en los tres planos de proyección. Obviamente, esta separabilidad disminuirá drásticamente al aumentar el número de muestras utilizadas.

En cada uno de los planos representados aparece el nombre del formante correspondiente a cada eje de coordenadas y el rango de valores en los que se encuentran las vocales. El resto de las figuras de este apartado seguirán la misma disposición, lo que sin duda facilitará la interpretación de los resultados.

La figura 5.34 muestra las posiciones de las vocales adyacentes a consonantes bilabiales. La extensión (variabilidad) que presenta cada vocal ha aumentado respecto al caso anterior, abarcándose en todas ellas frecuencias más bajas para cada uno de los formantes. Esta característica se ajusta perfectamente al comportamiento esperado, con todos los formantes descendiendo hacia un locus bajo forzado por las consonantes bilabiales.

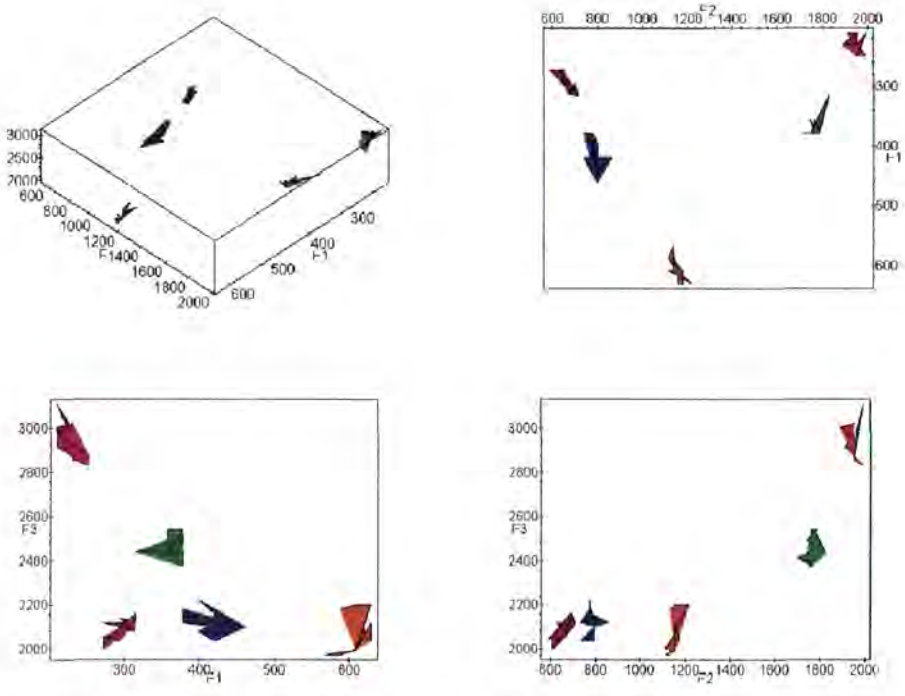


Figura 5.33

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales aisladas.

La figura 5.35 representa el grupo dental/interdental, en el que los valores frecuenciales a menudo son más altos que en el caso bilabial, esto se corresponde con la existencia de un locus de nivel medio. En la figura 5.36, correspondiente al grupo velar, la tendencia iniciada en el grupo anterior se confirma y amplía, presentándose altos valores frecuenciales de los formantes segundo y tercero que evolucionan hacia un locus alto característico de las consonantes velares. Así, por ejemplo, el segundo formante de la 'o' tiene dos ramas diferenciadas, una que muestra la posición estable de la vocal y la otra (hacia la derecha) que evoluciona hacia frecuencias superiores.

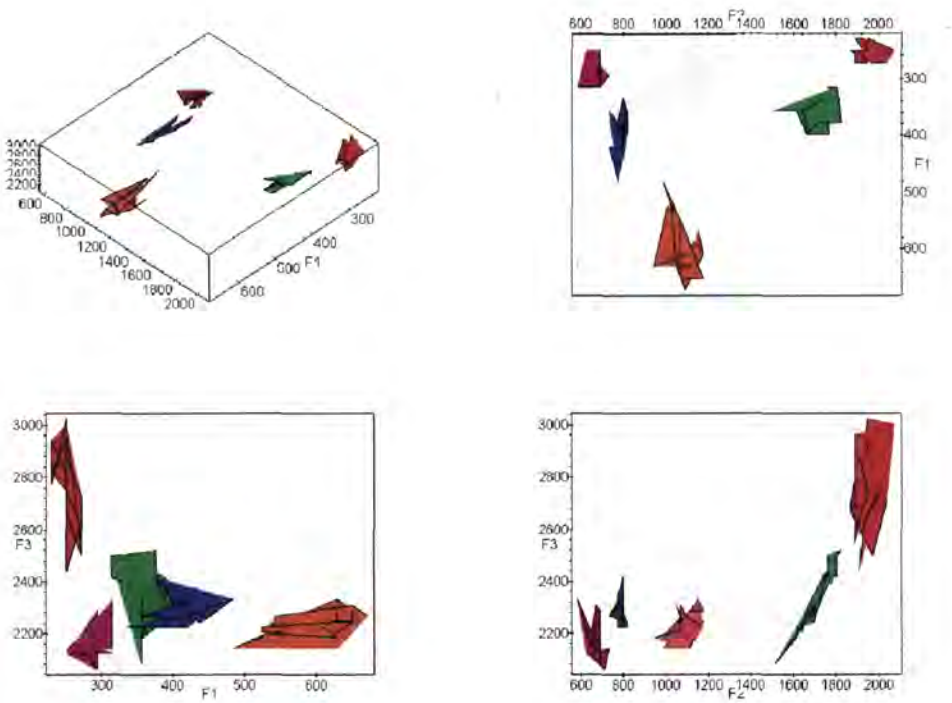


Figura 5.34

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales adyacentes a consonantes bilabiales.

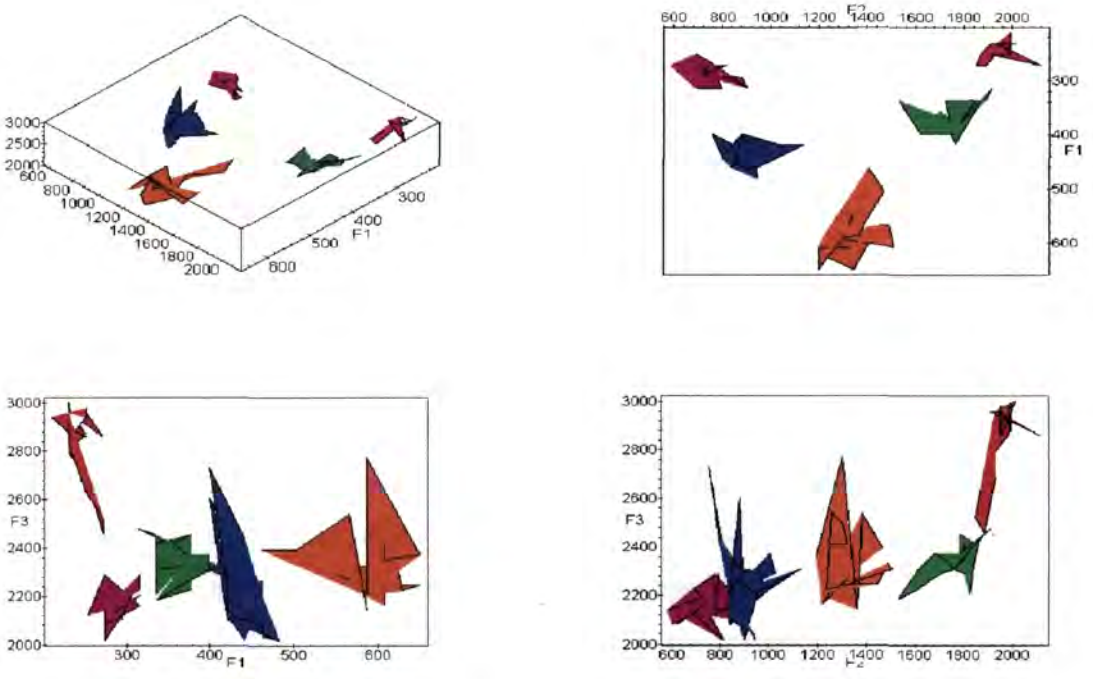


Figura 5.35

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales adyacentes a consonantes dentales/interdentales.

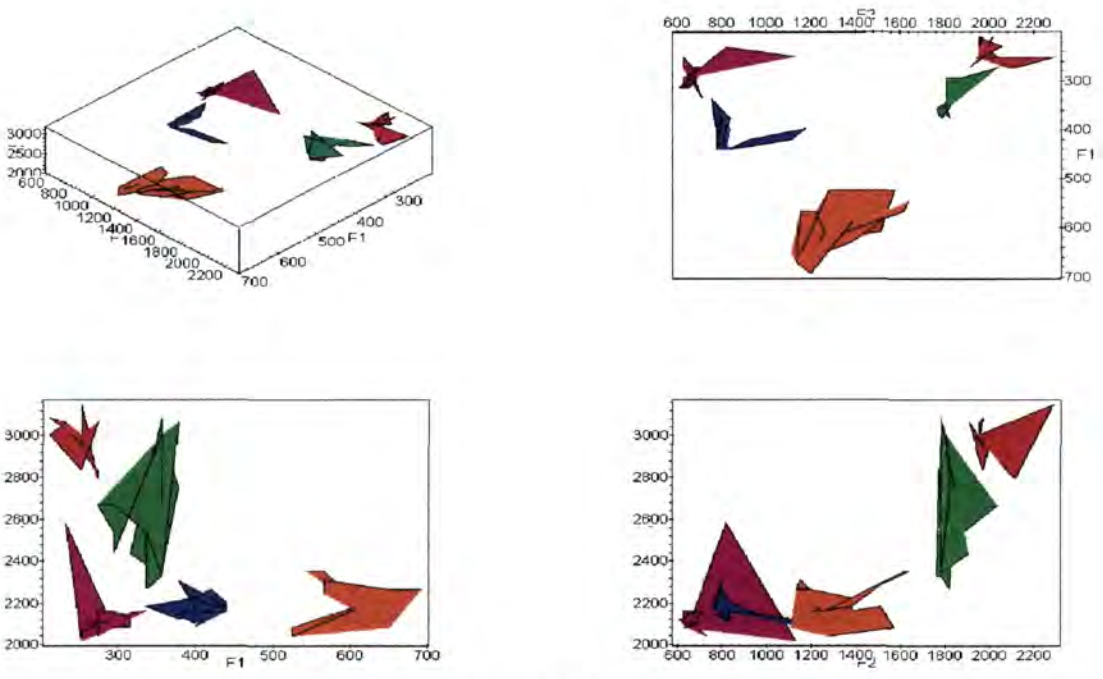


Figura 5.36

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales adyacentes a consonantes velares.

En la figura 5.37 se muestra la incidencia que presenta el punto de articulación de las consonantes en las posiciones frecuenciales que pueden adoptar las vocales. En esta figura se unifican los resultados de las cuatro anteriores y se representan en color amarillo las vocales aisladas, en rojo las vocales adyacentes a consonantes bilabiales, verde para las dentales/interdentales y azul para las palatales.

En la proyección F1-F2 de la figura 5.37, se aprecia la tendencia a la subida en frecuencias que adopta F2 a medida que aumenta la altura de los locus teóricos (bilabiales ⇒ menores frecuencias, velares/palatales ⇒ mayores frecuencias). También podemos establecer que las vocales aisladas (amarillo) adquieren un subconjunto muy reducido de las posibilidades de variación frecuencial que presentan las vocales ante distintos contextos de sonidos.

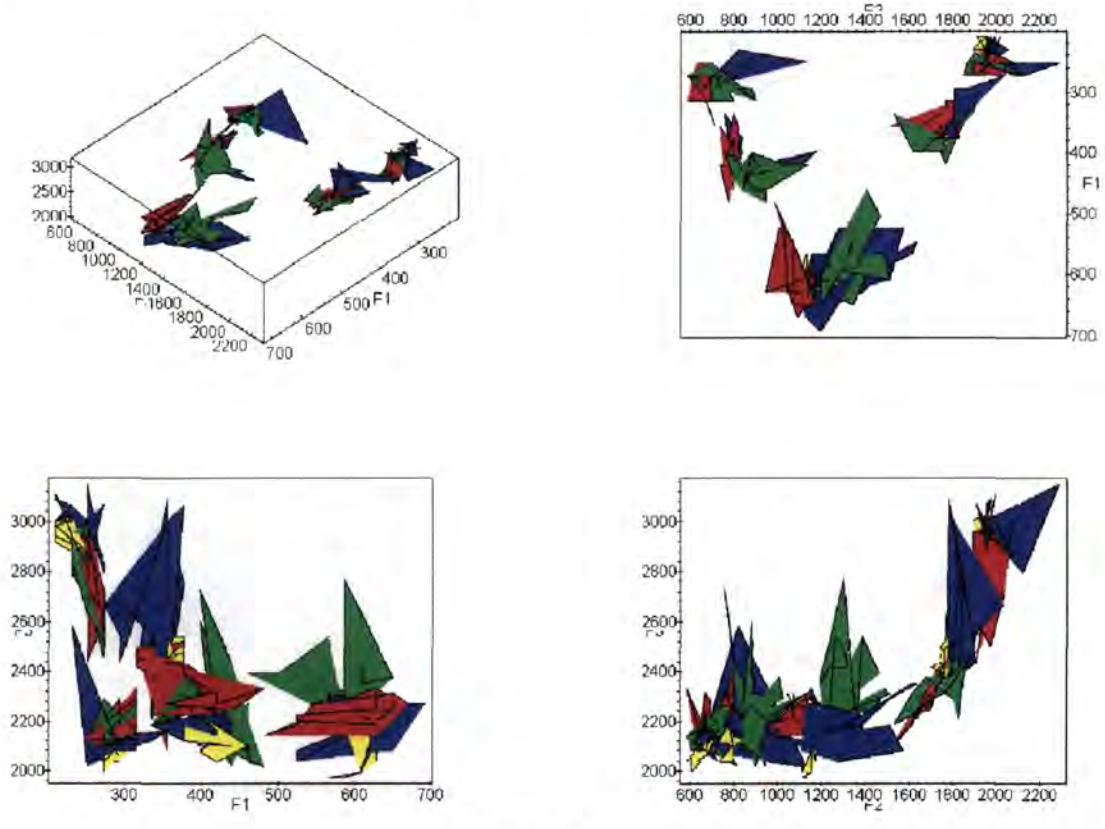


Figura 5.37

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en vocales aisladas (amarillo), vocales adyacentes a consonantes bilabiales (rojo), a consonantes dentales/interdentales (verde), a consonantes velares (azul).

En el plano F1-F3 se aprecian valores extremadamente altos en el tercer formante de las vocales, por ejemplo 3000 Hz en la 'e' ó 2600 Hz en la 'u', provocados por valores extremos situados muy cerca de las consonantes con las que se articulan los sonidos, también ocurre lo mismo con los valores bajos de las bilabiales, por ejemplo los 2500 Hz de la 'i'.

La figura 5.38 es la última que ilustra los resultados aislados del primer hablante, en ella se representan los valores que toman las vocales, sin tenerse en cuenta el grupo consonántico con las que se articulan. Todavía existe separabilidad lineal, aunque en este caso sólo en el plano F1-F2, que puede ser complementado con los otros dos.

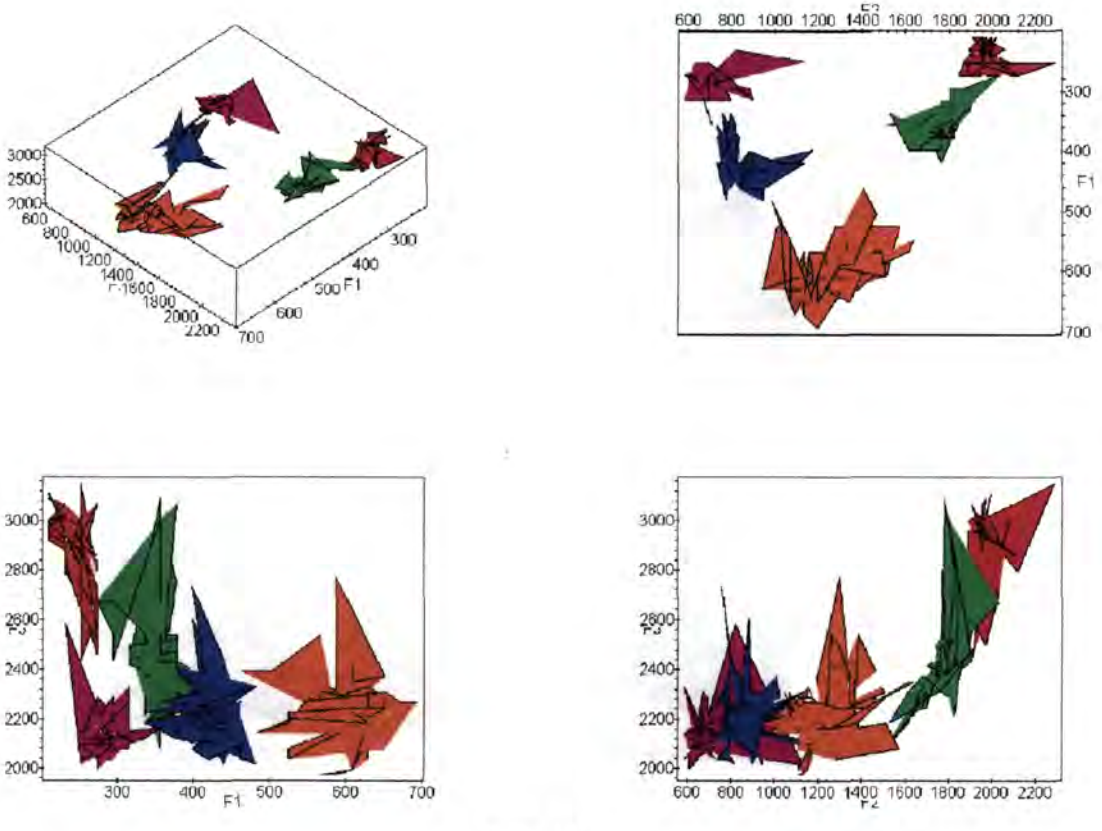


Figura 5.38

Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano.

i: rojo, e: verde, a: naranja, o: azul, u: magenta.

Las figuras 5.39 y 5.40 representan los resultados finales en el caso de un hablante femenino. Se confirman las diferentes posiciones de los formantes según esté situado el punto de

articulación de la consonante (figura 5.39), así como la separabilidad de vocales en el plano F1-F2 (figura 5.40).

Aunque la disposición de las vocales se mantiene frente al caso anterior, se puede observar que el rango en frecuencias de las escalas aumenta, concretamente F1 pasa de 700 Hz a 1000 Hz, y F2 de 2350 Hz a 2950 Hz, ambas aumentando el límite superior de la escala. F3 amplia su límite superior en 350 Hz y el inferior en 300 Hz

La fuerte subida en el rango de frecuencias se debe sin duda a la elevación del tono de voz producida en el cambio de un hablante masculino a otro femenino [BUS95]. Como es sabido, estas diferencias se pueden amortiguar presentando las distancias entre el tono fundamental y los formantes en lugar de las posiciones absolutas de los formantes en sí.

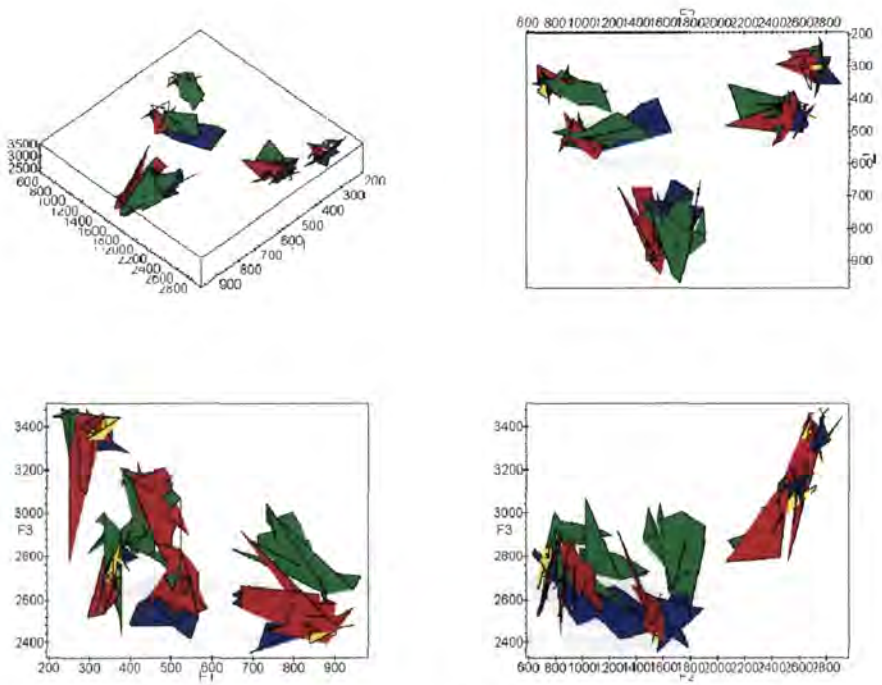


Figura 5.39

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos para un hablante femenino en vocales aisladas (amarillo), vocales adyacentes a consonantes bilabiales (rojo), a consonantes dentales/interdentales (verde), a consonantes velares (azul).

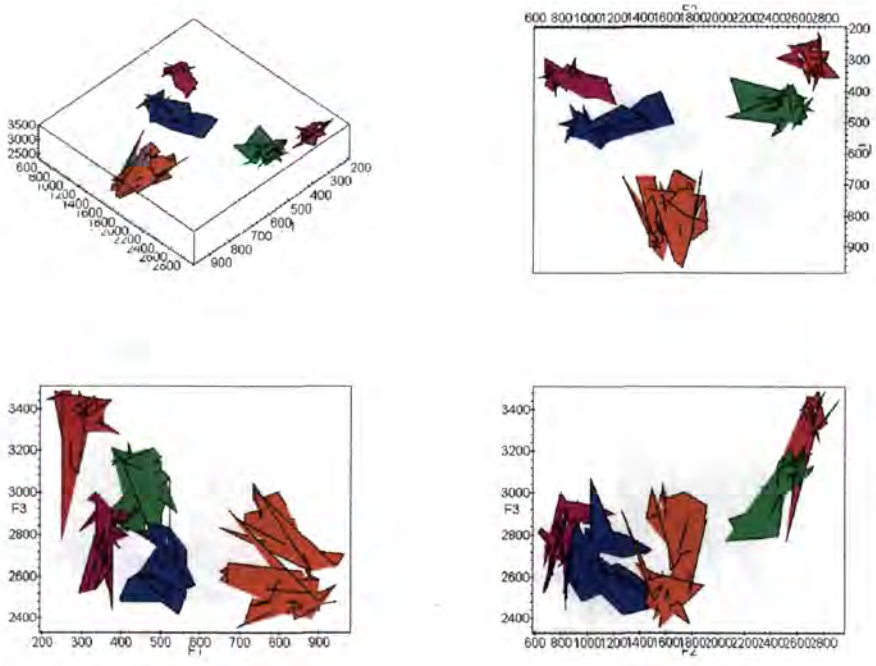


Figura 5.40
Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando un hablante femenino .
i: rojo, e: verde, a: naranja, o: azul, u: magenta.

Las figuras 5.41 y 5.42 ilustran los resultados del segundo hablante femenino tomado como referencia. Como se puede apreciar, la altura de sus formantes supera a las del caso anterior, y la disposición de las vocales se mantiene. La fuerte variación de la 'i' en el caso aislado (amarillo), aunque ha sido respetada, se debe a la influencia que ejerce la 'e' (contigua) por su proximidad en el tiempo en el momento de pronunciarse ambas una detrás de la otra.

Llegados a este punto, resulta importante resaltar las fuertes diferencias que existen entre distintos hablantes cuando se compara el rango de frecuencias de sus vocales [PET52], [HIL95] (comparar las figuras 5.42 y 5.38), a la vez que se mantienen sus posiciones relativas en los diferentes planos de formantes.

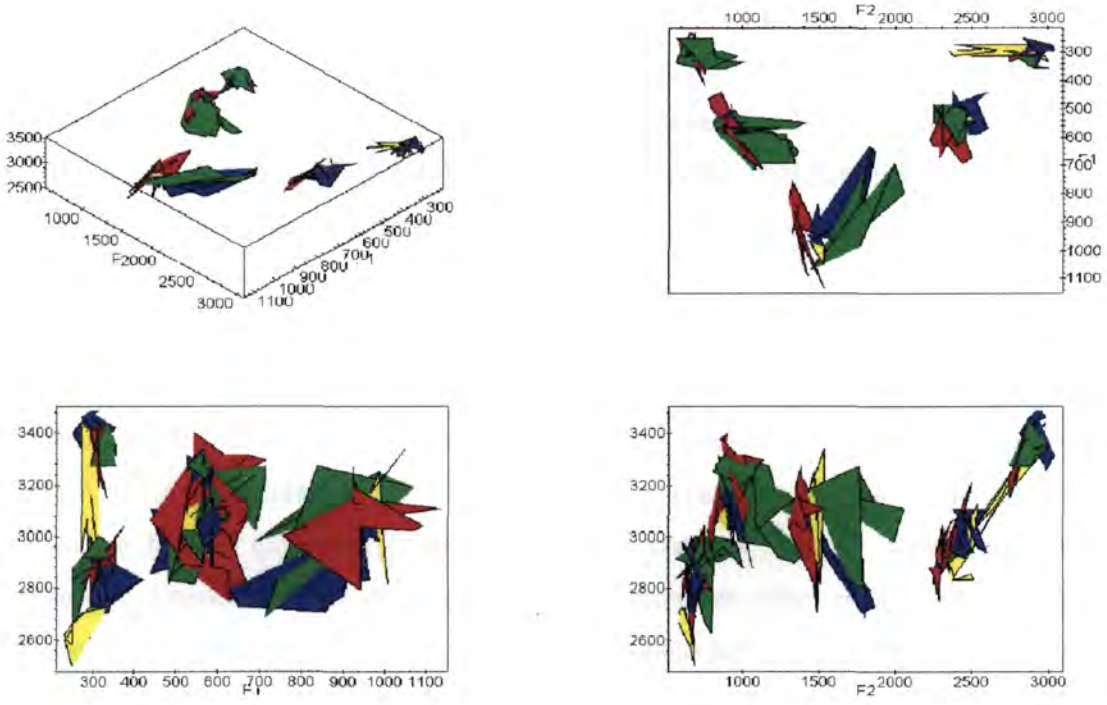


Figura 5.41

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos para un hablante femenino en vocales aisladas (amarillo), vocales adyacentes a consonantes bilabiales (rojo), a consonantes dentales/interdentales (verde), a consonantes velares (azul).

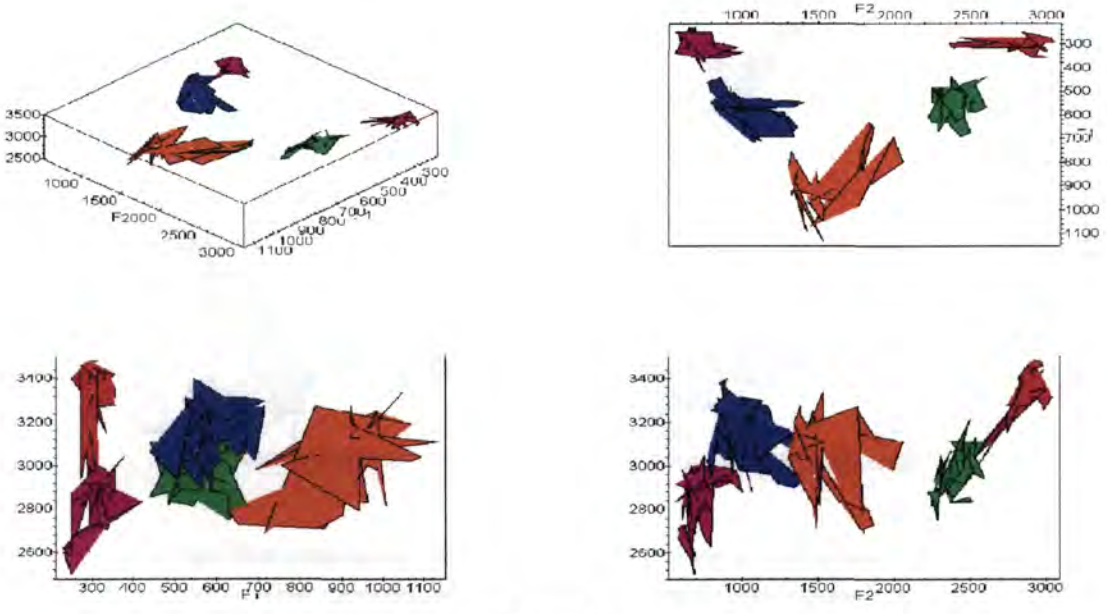


Figura 5.42

Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando un hablante femenino .

i: rojo, e: verde, a: naranja, o: azul, u: magenta.

Las figuras 5.43 y 5.44 presentan los resultados finales correspondientes al último hablante, en este caso masculino. Los principios fundamentales explicados en los ejemplos anteriores se mantienen, con lo que al conjuntar los resultados de varios hablantes podremos esperar el cumplimiento de los principios básicos esperados (rangos de frecuencias, evoluciones hacia los locus, etc.).

Para comprender algo mejor como los gráficos representados expresan la posición (y en parte la evolución) de los formantes, tomemos como ejemplo y referencia la vocal 'a' reflejada en el plano F1-F2 de la figura 5.43. En el caso de que la vocal sea anterior o posterior a una consonante bilabial (color rojo), tanto el primero como el segundo formante bajan en frecuencias. Cuando la vocal es adyacente a una consonante velar (color azul), F1 baja en frecuencias, mientras que F2 sube. En el caso dental/interdental (color verde) los formantes están bastante estables, con tendencia a bajadas en F1 y subidas en F2.

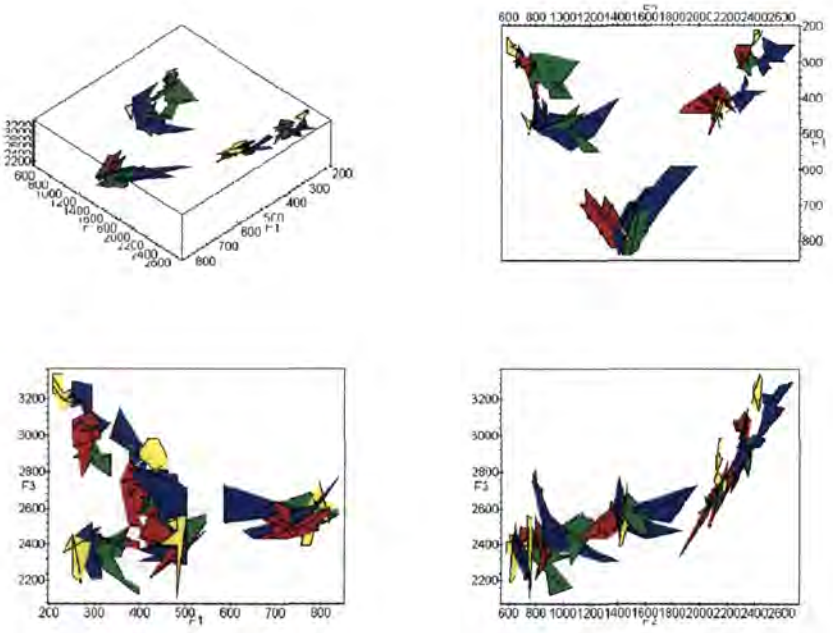


Figura 5.43

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos para un hablante masculino en vocales aisladas (amarillo), vocales adyacentes a consonantes bilabiales (rojo), a consonantes dentales/interdentales (verde), a consonantes velares (azul).

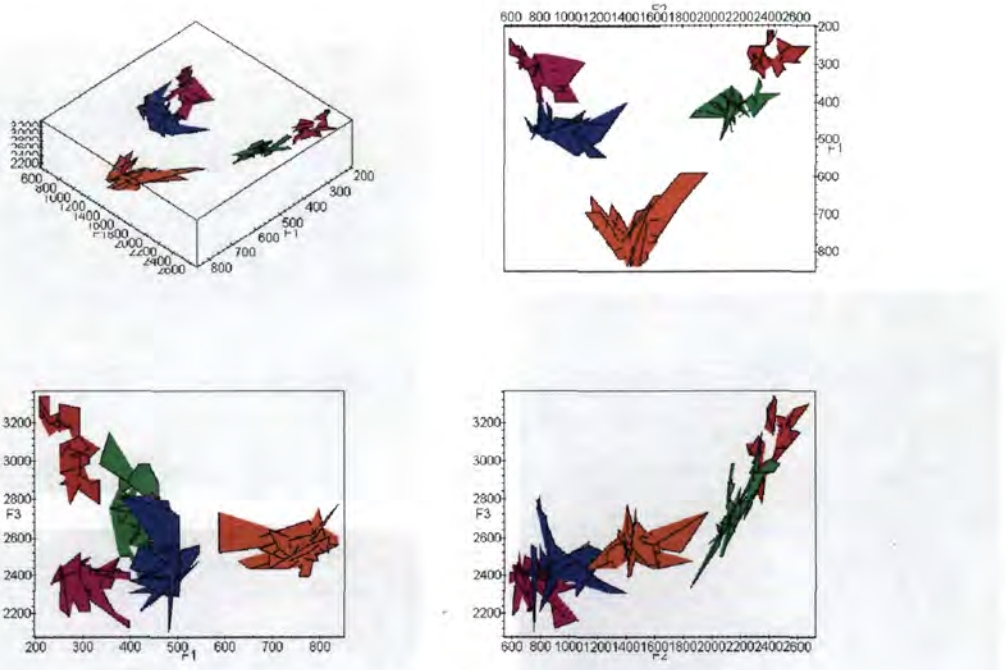


Figura 5.44

Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando un hablante masculino .

i: rojo, e: verde, a: naranja, o: azul, u: magenta.

Hay que aclarar que estos resultados no distinguen entre sonidos dentro de un grupo (por ejemplo, no se distingue entre [m], [p] y [β]), ni tampoco se hacen promedios de resultados, sino que éstos únicamente se recogen y presentan, con lo que podría ocurrir que el comportamiento mencionado en el párrafo anterior fuera cierto en el caso bilabial para [p], mientras que los sonidos [β] y [m] mantuvieran sus formantes estables.

La figura 5.45 presenta los espectros con formantes realzados de los que se han tomado los datos explicados en los párrafos anteriores ('a' del plano F1-F2 de la figura 12). El espectro situado en la parte superior izquierda de la figura se corresponde con los sonidos 'apa aβa ama', nos encontramos en el caso bilabial y podemos observar como en las tres secuencias, F1 y F2 descienden hacia un locus bajo, tal y como habíamos determinado.

El espectro situado en la parte inferior izquierda de la figura, contiene los sonidos 'aka aɣa aŋa'. Tal y como habíamos deducido, los primeros formantes presentan fuertes bajadas y los segundos fuertes subidas (salvo en las vocales adyacentes al sonido [ɣ]). Por último, en el espectro de la derecha se presentan los sonidos 'ata ada ama' que mantienen bajadas en F1 y subidas en F2, ambas de carácter moderado.

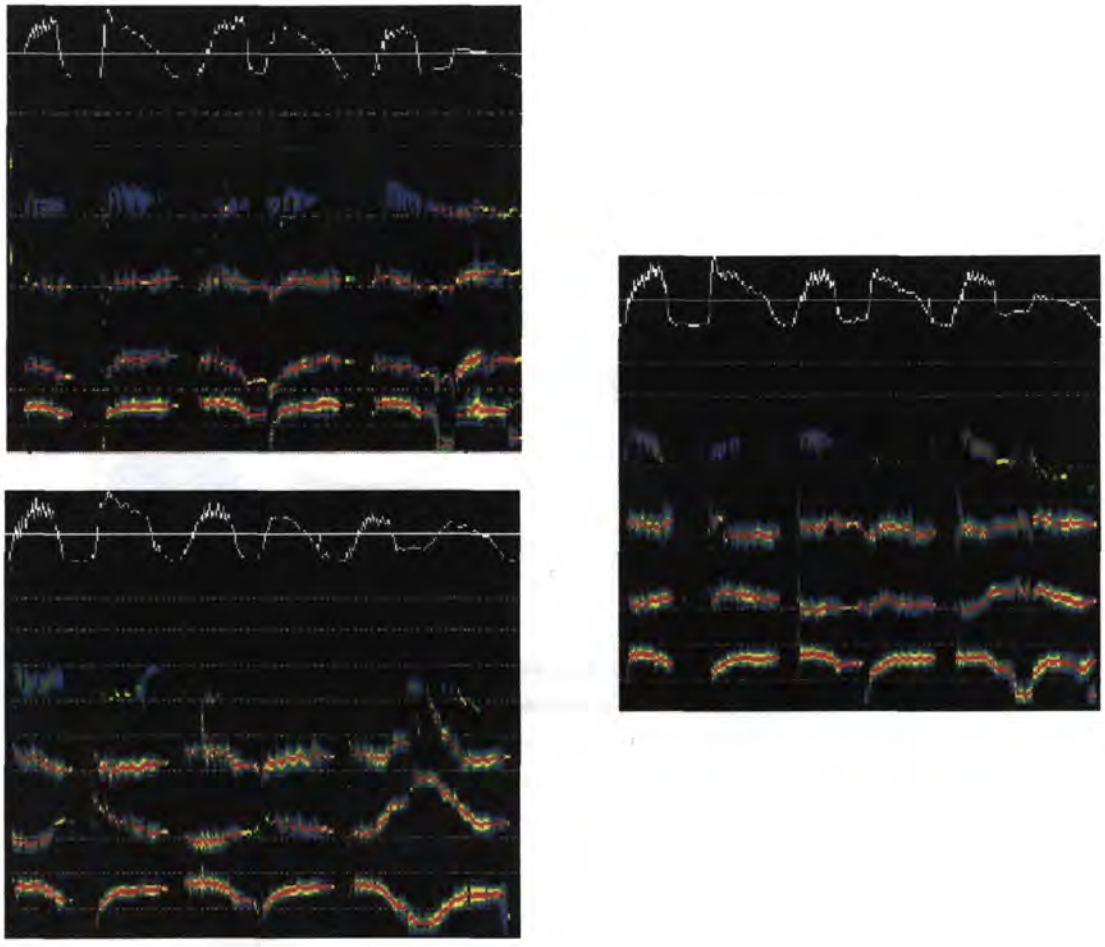


Figura 5.45
Espectros origen de los valores obtenidos en los formantes de la vocal 'a' en la figura 5.43. Espectro superior: 'apa aβa ama'. Espectro inferior: 'aka aɣa aŋa'. Espectro de la derecha: 'ata ad.a ana'.

En las figuras 5.46, 5.47 y 5.48 se agrupan los resultados pertenecientes a los dos hablantes masculinos. La figura 5.46, al igual que en cada caso individual, diferencia las vocales según el grupo de sus consonantes adyacentes. Las principales conclusiones que podemos obtener son:

1.- Las frecuencias que presentan los formantes son menores que en los casos femeninos.
2.- Se confirman las tendencias en las posiciones de los dos primeros formantes que se han comentado anteriormente y que dependen del punto de articulación de las consonantes vecinas.
3.- Las tendencias del tercer formante no se relacionan claramente con el punto de articulación.

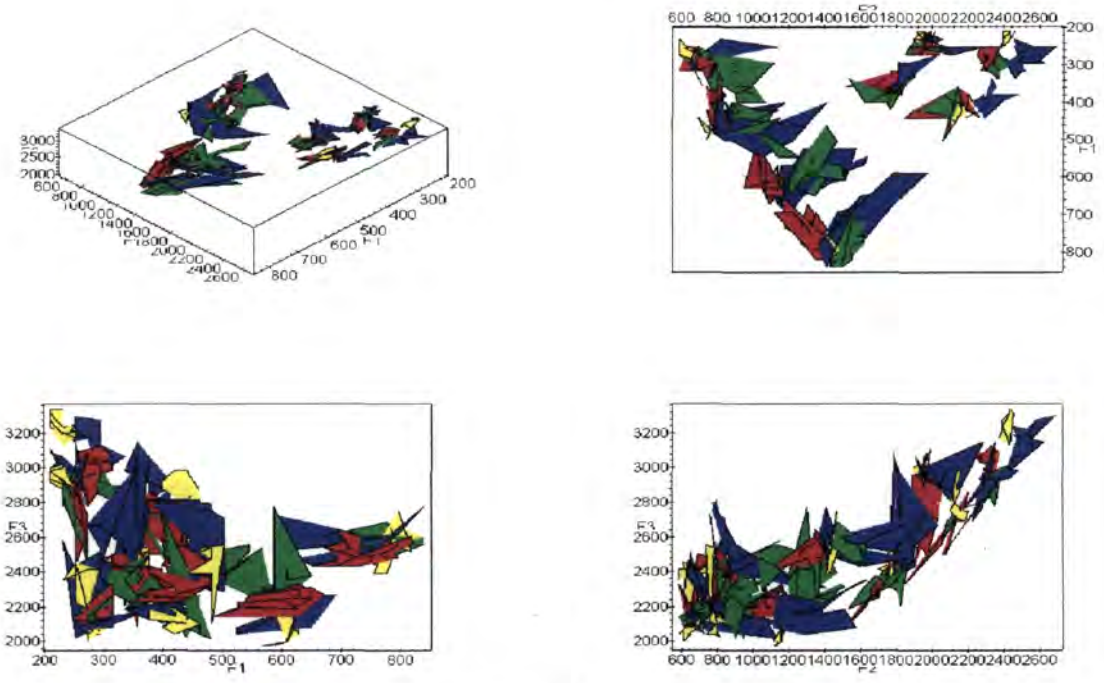


Figura 5.46

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en dos hablantes masculinos sobre vocales aisladas (amarillo), vocales adyacentes a consonantes bilabiales (rojo), a consonantes dentales/interdentales (verde), a consonantes velares (azul).

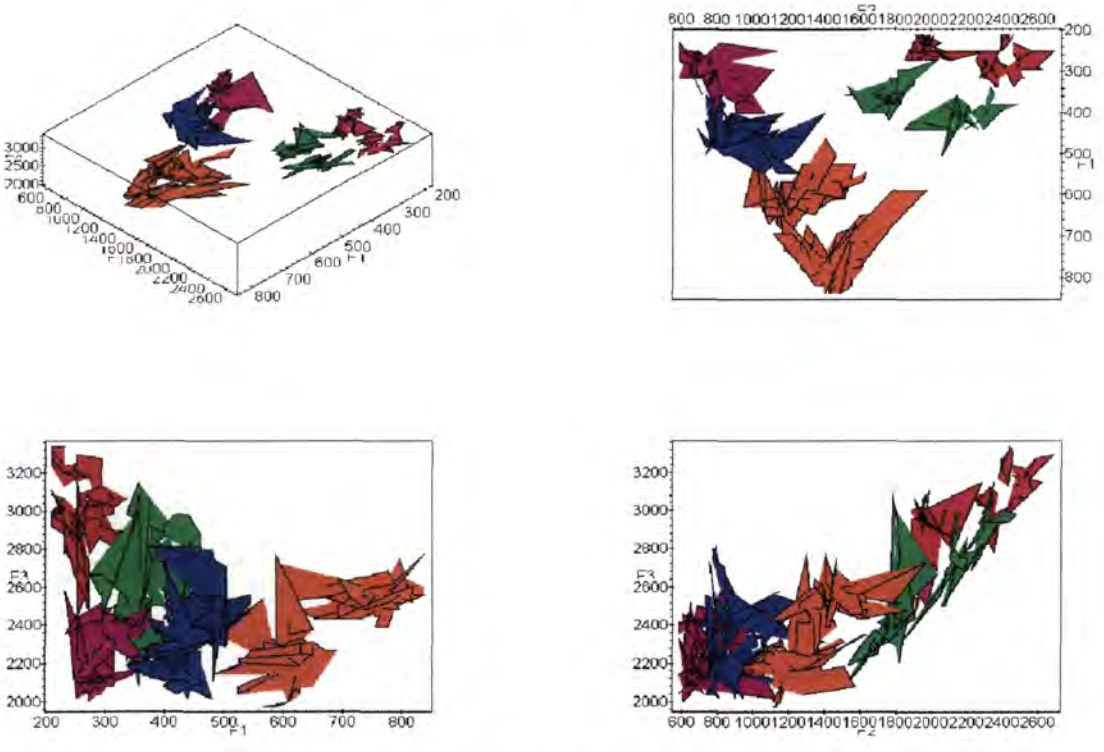


Figura 5.47

Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando dos hablantes masculinos .

i: rojo, e: verde, a: naranja, o: azul, u: magenta.

En la figura 5.47, los resultados se agrupan por vocales. Se puede apreciar como, aunque la separabilidad no se pierde en el plano F1-F2, las posiciones de cada vocal pueden variar bastante al introducirse diferentes hablantes. De nuevo, los planos F1-F3 y F2-F3 pueden ayudar a realizar identificaciones en caso de dudas en el plano F1-F2.

La figura 5.48 representa cada hablante con un color; en este caso, el hablante identificado con el color amarillo presenta menores frecuencias en los tres formantes que el hablante identificado por el color rojo. Otra característica importante es la gran diferencia que existe en las posiciones de las vocales 'i', 'e', 'a', que denotan una articulación muy diferente entre estos dos hablantes en las posiciones no abocinadas de la boca.

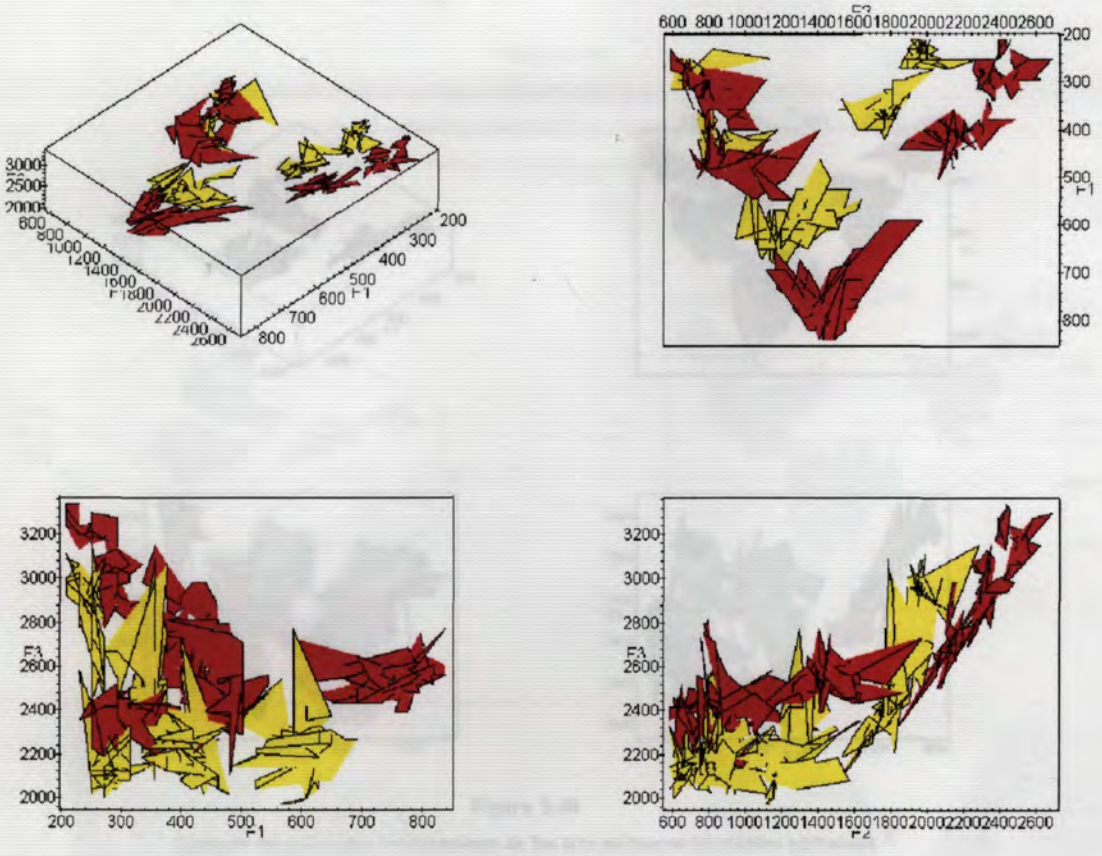


Figura 5.48

Posiciones de los formantes obtenidos en los dos hablantes masculinos.

Rojo: hablante masculino 1. Amarillo: hablante masculino 2.

Las figuras 5.49, 5.50 y 5.51 tienen exactamente el mismo significado que las anteriores, pero referidas a los dos hablantes femeninos. Respecto a la figura 5.49, cabe destacar una fuerte variación de las vocales adyacentes a consonantes dentales/interdentales, probablemente debido a la mayor altura de los formantes respecto a los locus, que fuerza la aparición de evoluciones con pendientes más pronunciadas.

La separabilidad de las vocales se mantiene en la figura 5.50, mientras que las diferencias de frecuencias entre hablantes (figura 5.51) se han reducido considerablemente frente al caso masculino. El hablante femenino de color verde tiene frecuencias más altas en todos los formantes que el representado por el color azul.

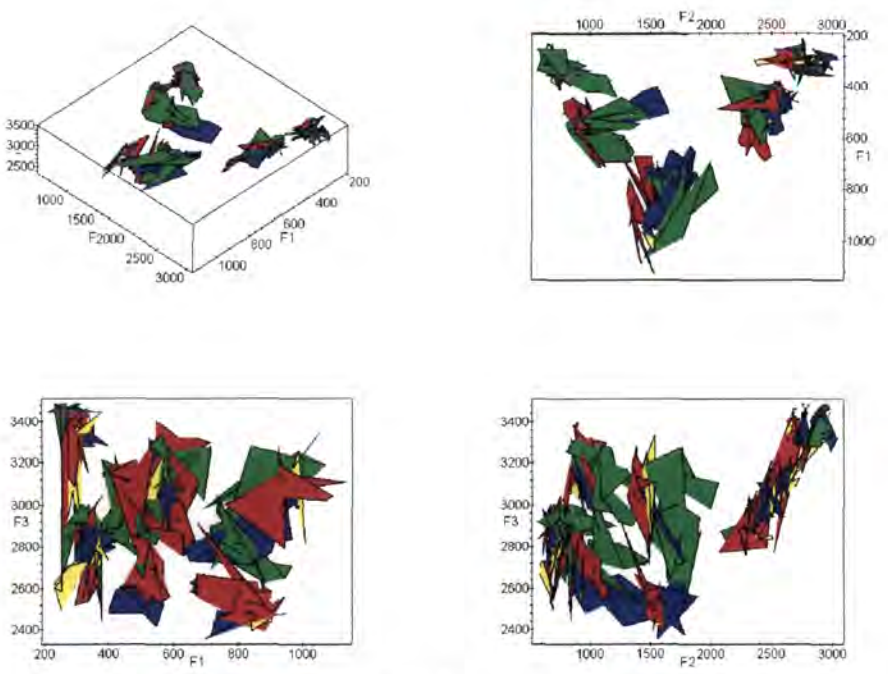


Figura 5.49

Ejemplo de posiciones frecuenciales de los tres primeros formantes obtenidos en dos hablantes femeninos sobre vocales aisladas (amarillo), vocales adyacentes a consonantes bilabiales (rojo), a consonantes dentales/interdentales (verde), a consonantes velares (azul).

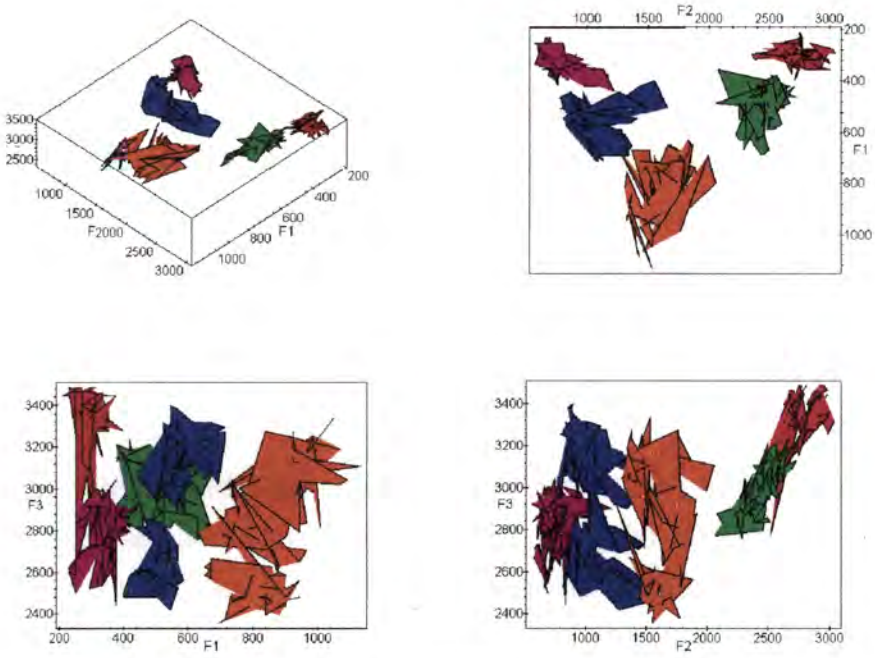


Figura 5.50

Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando dos hablantes femeninos .

i: rojo, e: verde, a: naranja, o: azul, u: magenta.

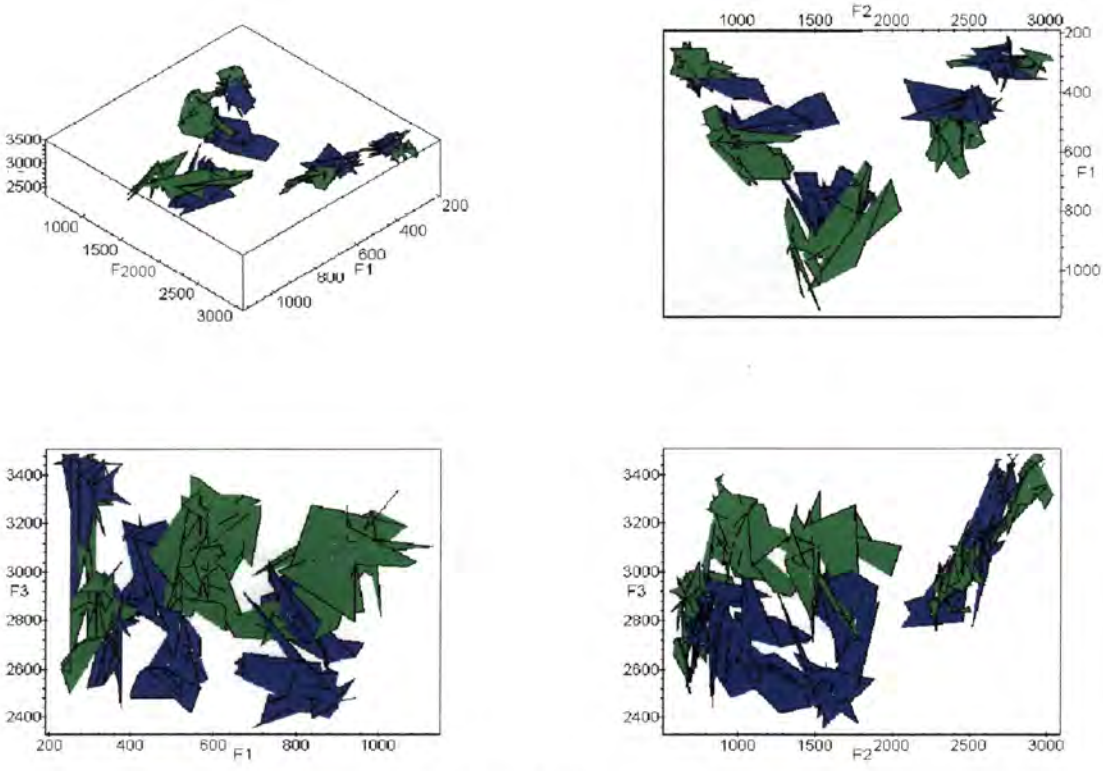


Figura 5.51

Posiciones de los formantes obtenidos en los dos hablantes femeninos.

Verde: hablante femenino 1. Azul: hablante femenino 2.

La figura 5.52 presenta la posición de las vocales unificando todos los resultados de las etapas anteriores. La separabilidad ha disminuido debido a la variedad de las muestras obtenidas y sobre todo a la utilización de varios hablantes; sin embargo, todavía es posible distinguir con facilidad la mayor parte de los casos. Los planos en los que interviene el tercer formante siguen ayudando a la identificación vocálica.

En la figura 5.53 se presentan las vocales diferenciadas por hablantes. Al haberse unificado los diferentes rangos de las escalas que abarca cada hablante por separado, el resultado ha sido que en los casos masculinos (con frecuencias más bajas) las vocales se han trasladado (comprimido) hacia los orígenes de frecuencias de cada plano representado.

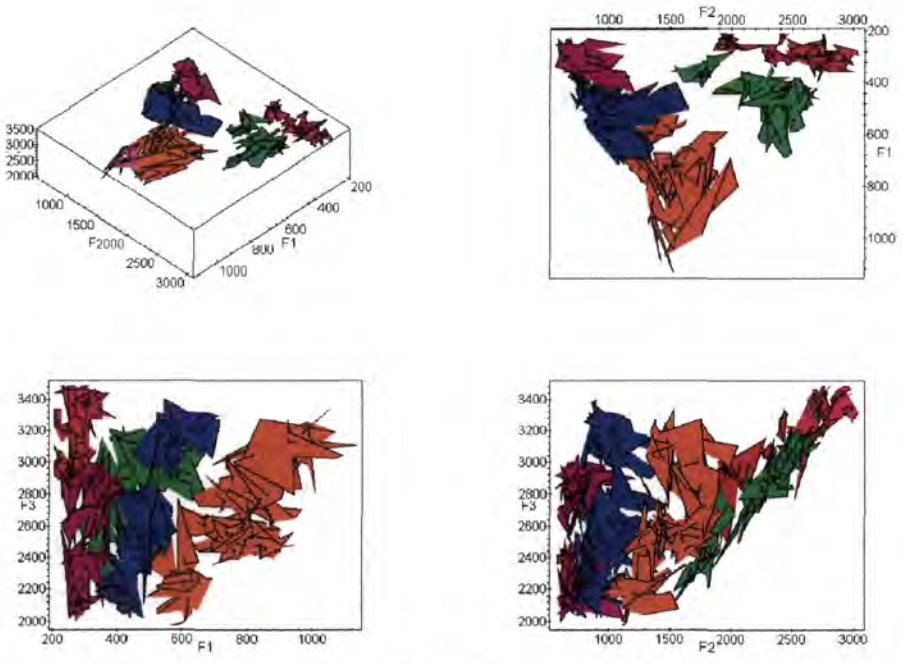


Figura 5.52

Ejemplo de posiciones frecuenciales de los tres primeros formantes en las vocales del castellano empleando a los 4 hablantes (2 masculinos y 2 femeninos). i: rojo, e: verde, a: naranja, o: azul, u: magenta.

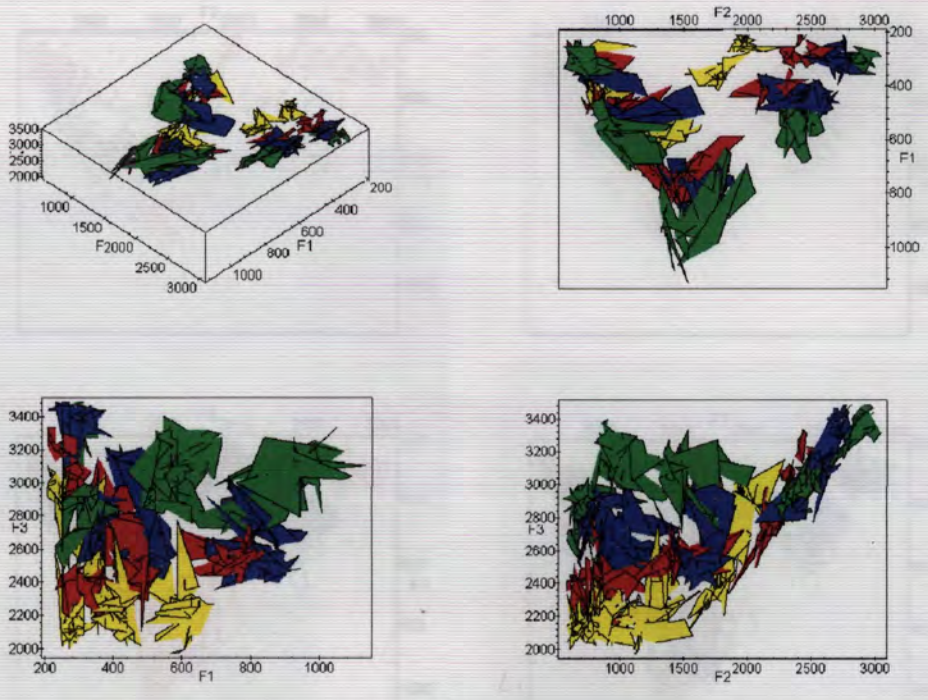


Figura 5.53

Posiciones de los formantes obtenidos empleando los 4 hablantes (2 masculinos y 2 femeninos). Verde: hablante femenino 1. Azul : hablante femenino 2. Rojo : hablante masculino 1. Amarillo : hablante masculino 2.

El fenómeno expresado en el párrafo anterior, explica como es posible que, por ejemplo, la 'a' del hablante 'amarillo' se confunda con la 'o' de los demás. Sencillamente estamos comparando valores absolutos de frecuencias, cuando en realidad lo que nos interesa son sus posiciones relativas (figuras 5.38, 5.40, 5.42 y 5.44).

La figura 5.53 ilustra muy claramente la incidencia del tono de la voz del hablante en las posiciones absolutas de sus formantes. La relación es directa e incide en F1, F2 y F3. la solución a este problema consiste en aplicar un sencillo factor de corrección a los valores frecuenciales obtenidos si conocemos a priori las características de cada hablante, en caso contrario, este factor lo obtendríamos de una correcta estimación de la frecuencia fundamental del hablante.

La figura 5.54 presenta los planos F1-F2 correspondientes a las figuras 5.52 y 5.53, pero con características de visualización diferentes: puntos y contornos, esto permite hacerse una idea aproximada de cual es la posición media de los formantes, que por ejemplo, en el caso de la 'o', normalmente no tiene frecuencias tan altas en F2 como parecía en la figura 5.52.

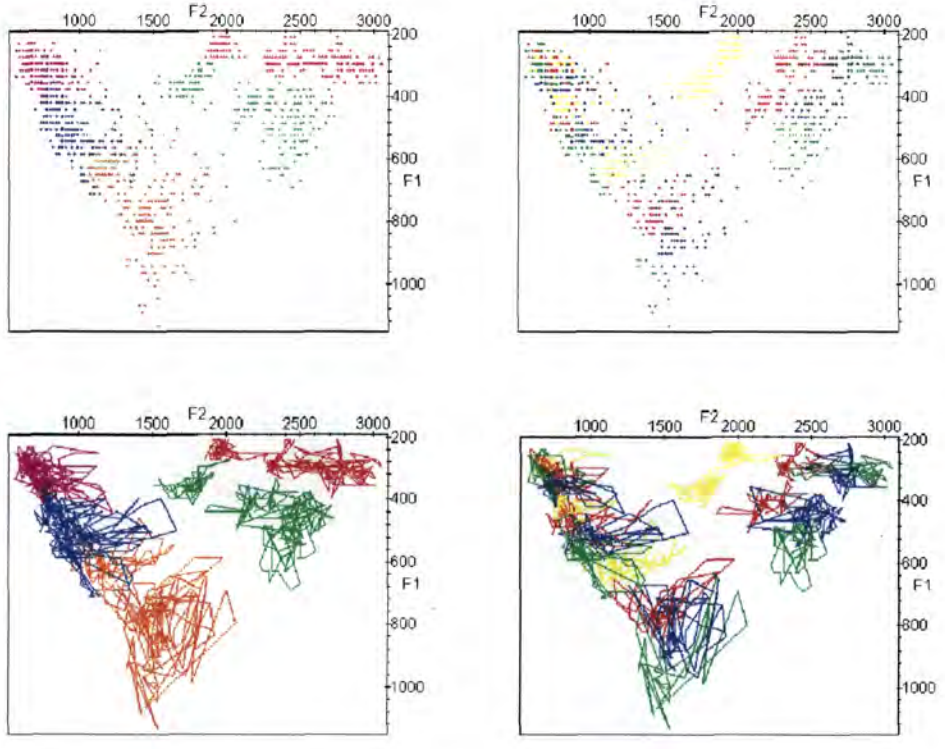


Figura 5.54

Planos F1-F2 correspondientes a las figuras 5.52 y 5.53 presentados como nube de puntos y por superficies transparentes.

La figura 5.55 contiene los espectros de voz de la secuencia 'eβe ed.e eγe' pronunciada por tres hablantes diferentes. El gráfico de la parte inferior izquierda corresponde a uno de los casos femeninos, mientras que los otros dos espectros han sido generados por cada uno de los hablantes masculinos grabados en este estudio. Comparando las posiciones de los formantes, se puede observar las diferencias en las alturas de las frecuencias detectadas a lo largo de este apartado.

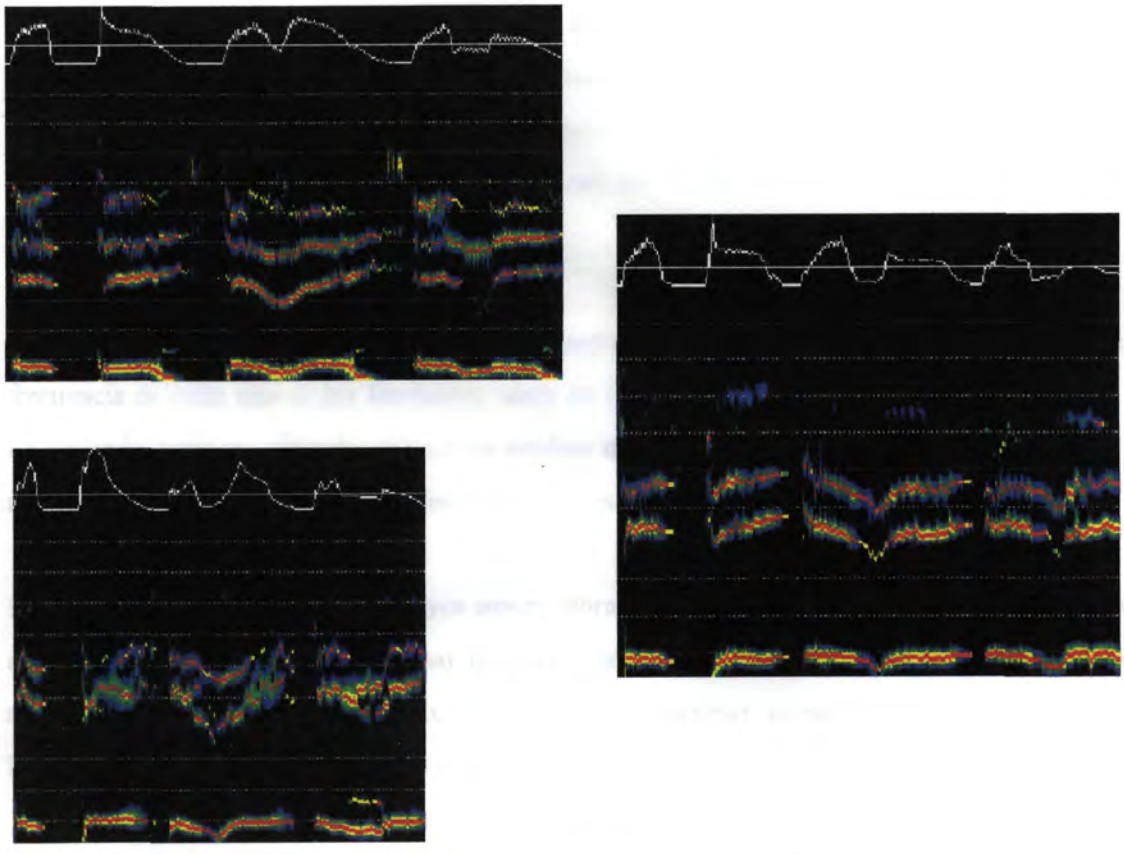


Figura 5.55

Secuencia 'eβe ed.e eγe' pronunciada por tres hablantes diferentes.

5.5.4 ESTADÍSTICOS RELEVANTES

Con los resultados obtenidos en el apartado anterior, nos podemos hacer una idea de la situación de los formantes para las distintas vocales y hablantes empleados en el estudio. A la vista de los gráficos presentados también se intuyen las posiciones medias de los formantes, sin embargo, no es posible deducir las distribuciones de frecuencias más representativas de los datos representados.

El primer objetivo que se persigue en este punto es la obtención de las distribuciones de frecuencia de cada uno de los formantes, tanto de forma aislada como agrupados por vocales. En segundo lugar se pretende realizar un análisis discriminante con la intención de conseguir una mayor separabilidad entre vocales y/o hablantes que la hallada en el apartado anterior.

El planteamiento de objetivos de mayor envergadura no resulta adecuado en este estudio debido a que para ello sería necesario tomar un mayor número de muestras sobre un conjunto más amplio de hablantes, en nuestro caso, tal y como cabría esperar, no se cumplen adecuadamente las hipótesis básicas necesarias para realizar análisis de varianzas, correlaciones, etc.

En primer lugar se ha hallado la distribución de frecuencias de cada uno de los formantes para la totalidad de las vocales y los hablantes utilizados, los resultados obtenidos se presentan en la figura 5.56. En la gráfica superior correspondiente al primer formante, se aprecia con claridad como los valores más altos (comunes) se corresponden a las frecuencias espectrales más bajas, esto es así debido a la acumulación de los casos de la 'i' y la 'u' por un lado, y la 'e' y la 'a' por otro, ambos grupos con valores solapados en la posición del primer formante. Esta situación nos da una idea de las dificultades que nos podemos encontrar para conseguir aspectos tan básicos como la normalidad de los datos de entrada.

La segunda y tercera gráfica corresponden respectivamente al formante dos y tres. El segundo formante presenta una distribución de frecuencias que se centra en los valores espectrales más bajos, esto es así debido a que en este formante los mayores solapamientos frecuenciales se dan en las vocales 'o' y 'u', ambas con F2 pequeño. En el tercer formante, la distribución de frecuencias presenta un comportamiento (respecto a los casos anteriores) más independiente de las vocales de las que se toman las muestras. Los conceptos expuestos se pueden comprobar fácilmente analizando con detalle la figura 5.52 mostrada en el apartado anterior.

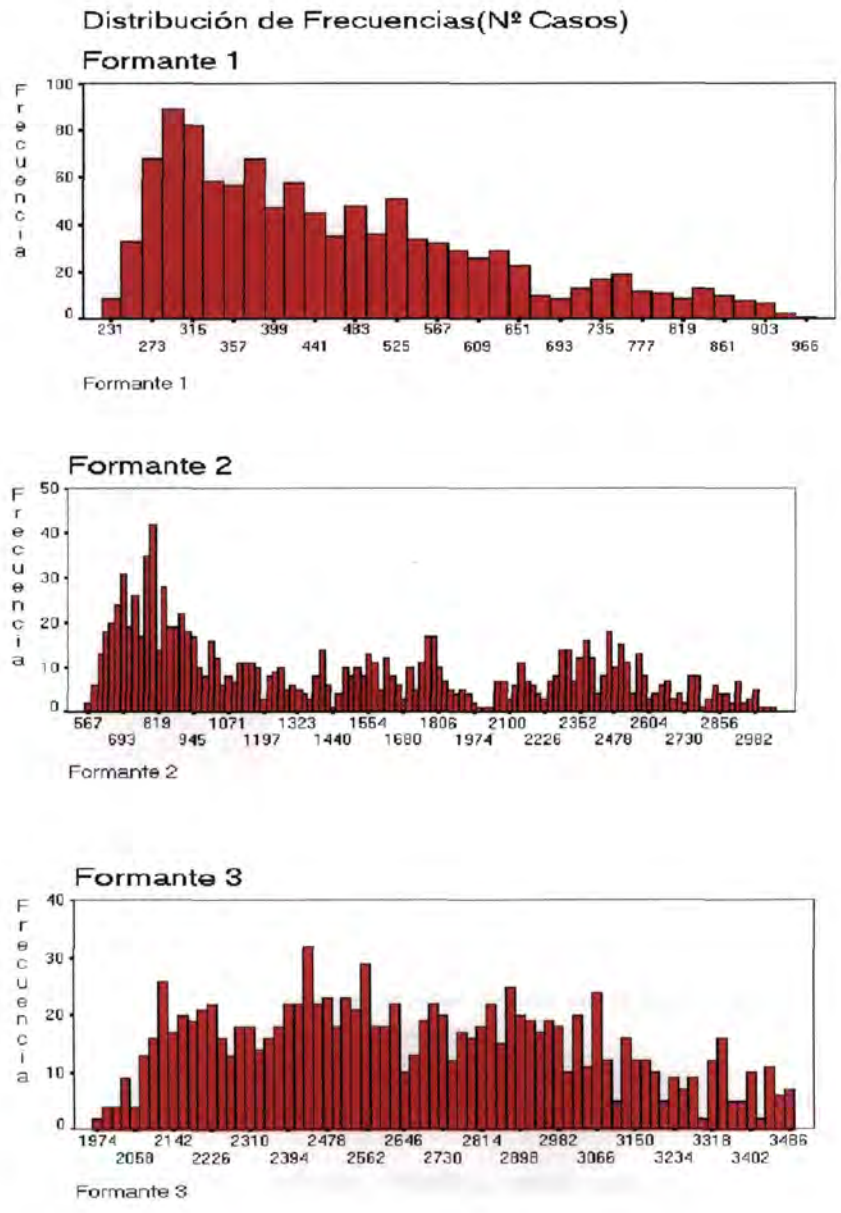


Figura 5.56

Distribución de frecuencias de cada uno de los formantes tomando los datos referentes a todas las vocales.

Las tres siguientes figuras, presentan para cada uno de los formantes la distribución de frecuencias que se da en las cinco vocales estudiadas. Si el estudio se hubiera realizado con un sólo hablante o con un gran número de ellos, las distribuciones se acercarían a la normal, sin embargo, al haberse utilizado únicamente cuatro personas y además con marcadas diferencias en las alturas espectrales entre los hombres y las mujeres, nos encontramos con distribuciones de frecuencias que nos indican la distancia espectral que existe entre hablantes o grupos de hablantes.

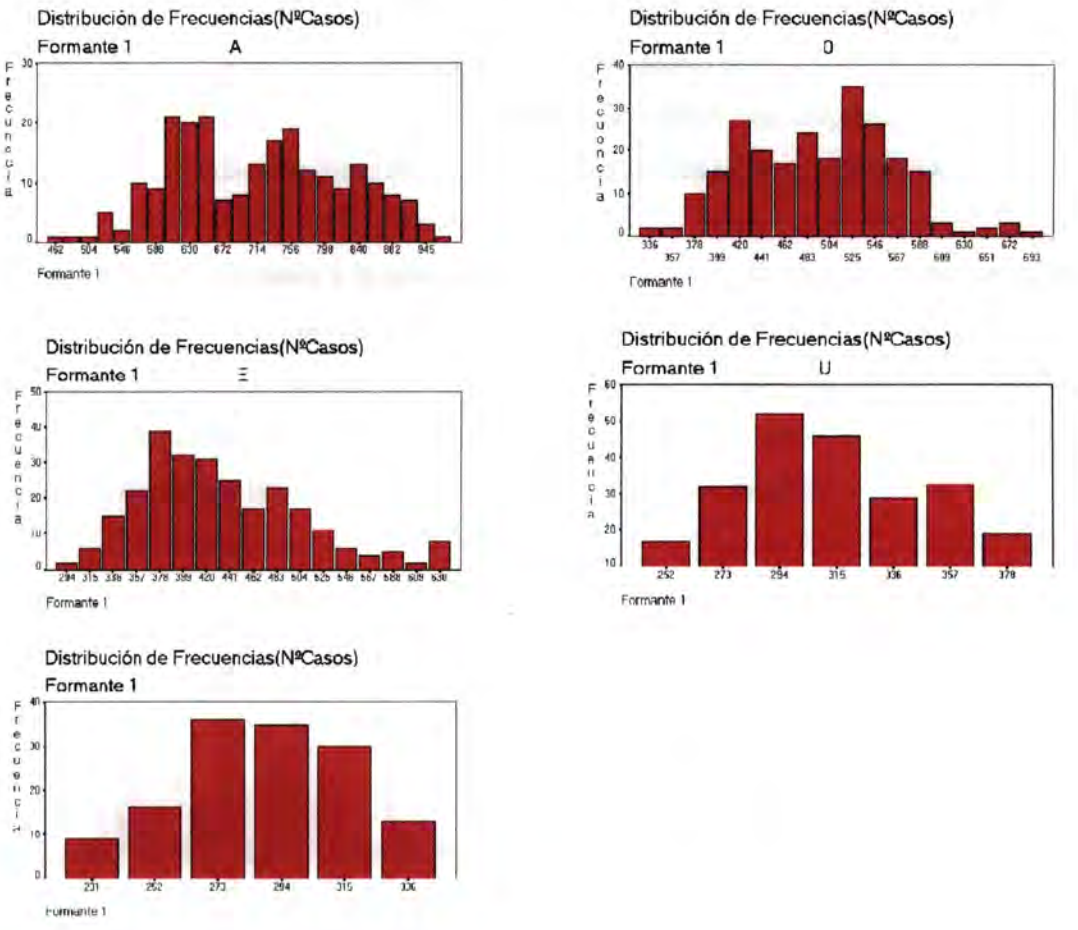


Figura 5.57

Distribución de frecuencias del primer formante para las cinco vocales estudiadas.

Para entender y comprobar los resultados obtenidos, resulta adecuado apoyarnos en la gráfica 5.53. Así, por ejemplo, la distribución de frecuencias del primer formante en [a], nos indica la existencia de dos casos diferenciados, uno centrado en los 630 Hz y otro (mas extenso) alrededor de los 790 Hercios. La confirmación de esta observación se encuentra a la vista de la gráfica del apartado anterior (5.53) donde se aprecia con claridad como el primer formante del hablante masculino representado con el color amarillo, se encuentra en frecuencias bajas centradas alrededor de los 600 Hz, mientras que en los demás hablantes (rojo, azul y verde) se agrupa en una amplia franja entre los 600 y los 1000 Hz con especial densidad entre los 750 y 850 Hz (figura 5.54), lo que se corresponde con la distribución de frecuencias mostrada.

Llegados a este punto, es importante destacar que para que razonamientos del tipo realizado en el párrafo anterior tengan validez, es necesario que el número de muestras tomadas para cada uno de los hablantes en cada una de las vocales sea similar, esta circunstancia debería ser forzada en estudios que pretendan dotar de universalidad a los resultados obtenidos.

La figura 5.58 es equivalente a la anterior, pero en este caso las distribuciones corresponden al segundo formante. Los comentarios realizados sobre la vocal 'a' del primer formante se pueden aplicar perfectamente al segundo formante de la vocal 'e'. También en este caso el hablante masculino representado con el color amarillo fuerza una separación en la distribución de frecuencias respecto a las otras tres personas involucradas en el estudio.

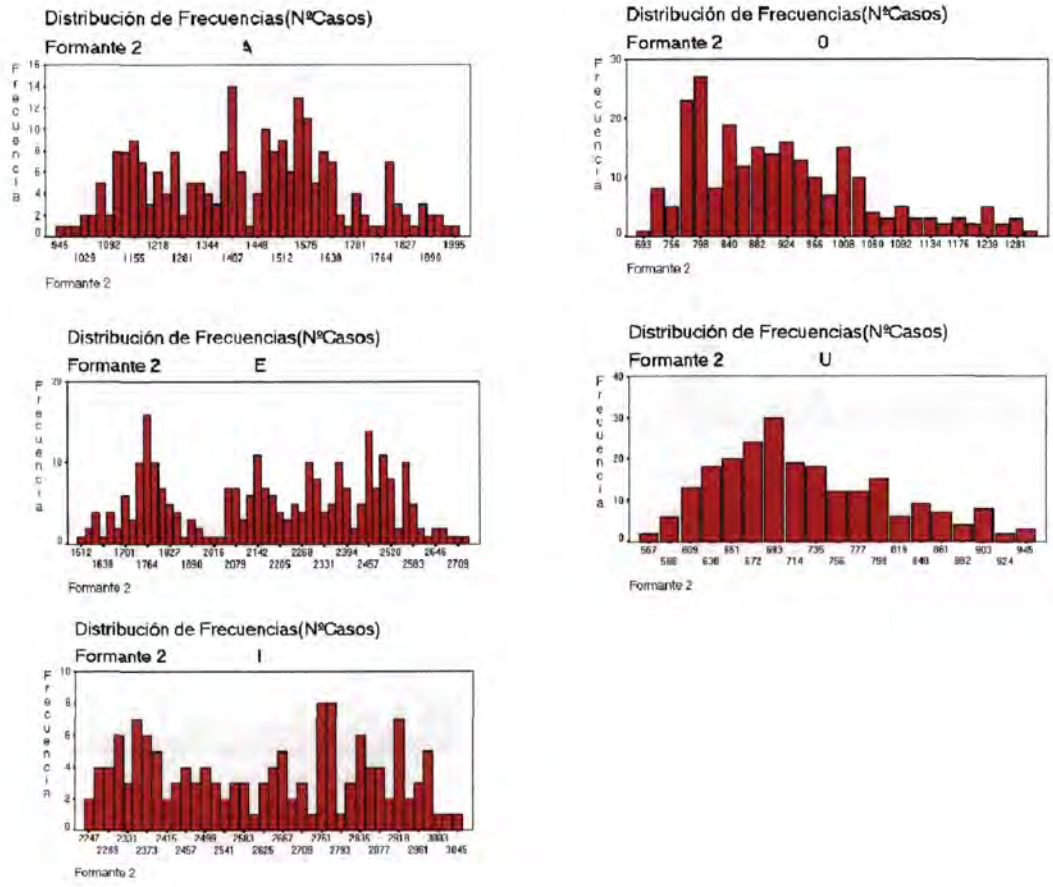


Figura 5.58

Distribución de frecuencias del segundo formante para las cinco vocales estudiadas.

La figura 5.59 correspondiente al tercer formante proporciona una ayuda mayor que las anteriores, puesto que la información equivalente que presentan las gráficas de las figuras 5.52 y 5.53 es difícil de obtener en el formante tres. En este caso, por ejemplo, en [u] se diferencian tres grupos, el primero centrado en 2100 Hz correspondiente al hablante 'amarillo', el segundo centrado en 2400 Hz correspondiente al 'rojo' y el tercero alrededor de los 2800 Hz correspondiente a la unión de los dos hablantes femeninos.

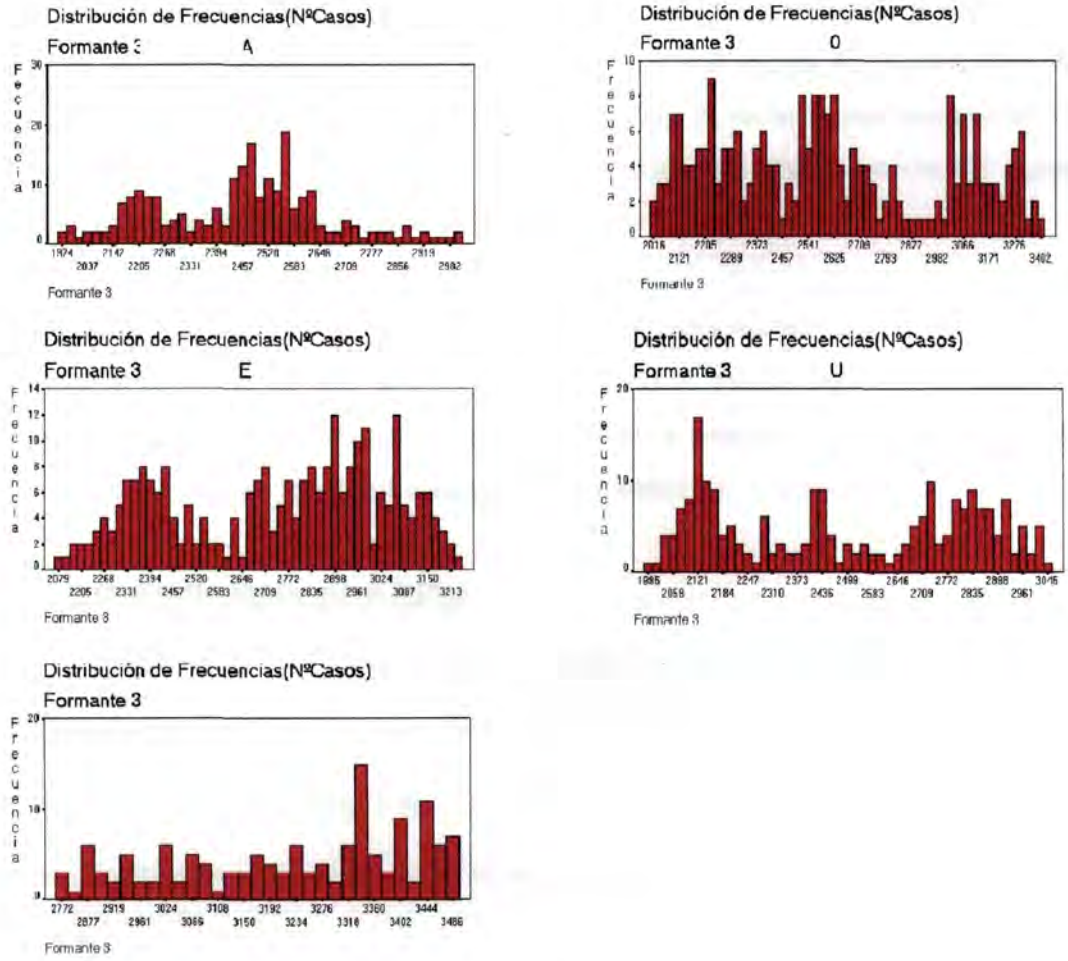


Figura 5.59
Distribución de frecuencias del tercer formante para las cinco vocales estudiadas.

Una última observación muy importante es que si solapamos las distribuciones de frecuencia de cada uno de los formantes, obtendremos gráficamente la probabilidad de que un formante

hallado en una frecuencia dada en un espectro se corresponda con el primero, segundo o tercero de alguna de las vocales (figura 5.56). Esta situación se hace más interesante teniendo en cuenta que podemos afinar estas probabilidades separando los casos por vocales (figuras 5.57 a 5.59).

Como aplicación de lo expuesto en el párrafo anterior, consideremos el caso de que nos encontramos un formante en el espectro situado en 815 Hercios, rodeado (frecuencialmente) de polos que no presentan una secuencialidad temporal clara. La pregunta que nos surge es a qué formante de qué vocal puede corresponder esta posición (815 Hercios). Observando la figura 5.56 podemos determinar que existe una pequeña probabilidad de ser un primer formante, una gran probabilidad de ser un segundo y una probabilidad nula de corresponderse con F3. Para conseguir una información más exacta acudimos a las distribuciones de frecuencia de cada vocal para cada formante. En la figura 5.57 deducimos que de ser un primer formante se trata de la vocal 'a'. Con la figura 5.58 se puede determinar que podría tratarse del segundo formante de una 'o' (bastante probable) o de una 'u'.

Con el fin de conseguir una mayor separabilidad (reconocimiento) de las vocales y los hablantes involucrados en este estudio, se ha acudido al uso de técnicas estadísticas de obtención de funciones discriminantes basadas en nuestros parámetros de partida (FR1, FR2, FR3). Los detalles de los resultados conseguidos se muestran en la siguiente información:

Funciones discriminates para letras

F1= -0.0130388 FR1 + 0.0038686034 FR2 + 0.00040029534 FR3 - 1.1588536
F2= 0.0102762 FR1 + 0.003274727 FR2 - 0.00293073743 FR3 - 1.7841424
F3= 0.00242425425 FR1 - 0.000766235747 FR2 + 0.0036240845 FR3 - 8.9421914

Se clasifica correctamente el 96.78 %

Correlación entre las puntuaciones de los valores obtenidos para cada función y el formante respectivo.

Función	Formante	Correlación
F1	FR1	-0.43076
F1	FR2	0.59472
F1	FR3	0.18834
F2	FR1	0.65431
F2	FR2	0.63975
F2	FR3	0.00729
F3	FR1	0.62155
F3	FR2	0.48686
F3	FR3	0.98208

De los datos se deduce en las funciones discriminantes por sonidos que la primera función (F1) participa especialmente de FR2 de forma directa y en menor medida de FR1 de forma inversa. La influencia de FR3 es baja. La segunda función se obtiene a partes iguales de FR1 y FR2, mientras que F3 se basa en los tres formantes, con especial influencia en el tercero.

El porcentaje de la clasificación es muy alto (96.78) teniendo en cuenta la sencillez del método lineal de obtención de las funciones discriminantes.

Funciones discriminantes para hablantes

F1= 0.00228376041 FR1 - 0.001002584407 FR2 + 0.00565281551 FR3 - 14.7045937
F2= -0.00296569947 FR1 + 0.00135440172 FR2 - 0.000563246434 FR3 + 0.7923895
F3= 0.0058234768 FR1 + 0.000933413241 FR2 - 0.00104506038 FR3 + 1.166026

Se clasifica correctamente el 63.93 %

Correlación entre funciones y formantes

Función	Formanate	Correlación
F1	FR1	0.10761
F1	FR2	0.16333
F1	FR3	0.79589
F2	FR1	-0.50762
F2	FR2	0.90356
F2	FR3	0.60543
F3	FR1	0.85483
F3	FR2	0.39610
F3	FR3	-0.00379

En este caso, el porcentaje de clasificación aunque no parece muy alto, resulta sorprendentemente bueno (con cuatro hablantes el azar nos brinda un 25%) si tenemos en cuenta que la altura de los formantes principales no es a priori un método adecuado para realizar reconocimiento de hablantes.

En la figura 5.60 se presentan tres gráficas con el mismo planteamiento y significado que las elaboradas en el apartado anterior, solo que en este caso, los ejes no representan valores frecuenciales de los formantes (FR1, FR2, FR3), sino posiciones de las funciones discriminantes halladas (F1, F2, F3). El resultado esperado es una mayor separación entre vocales que la visualizada en la figura 5.52.

La primera observación que se puede realizar, es que las diferencias entre hablantes quedan muy amortiguadas. En segundo lugar, la separación que existe entre las vocales en la primera gráfica es mejor que la hallada en el apartado anterior, no existiendo apenas intersecciones, lo que explica el 96.78% de aciertos en la clasificación. La función 3 se presenta inútil para la consecución del objetivo perseguido, puesto que todas las vocales adquieren posiciones similares cuando se aplica dicha función.

Como consecuencia de los resultados obtenidos, parece conveniente aplicar este sencillo paso de obtención de funciones discriminantes cuando se pretende realizar una clasificación vocálica lineal atendiendo a las posiciones espectrales de los formantes como parámetros de entrada.

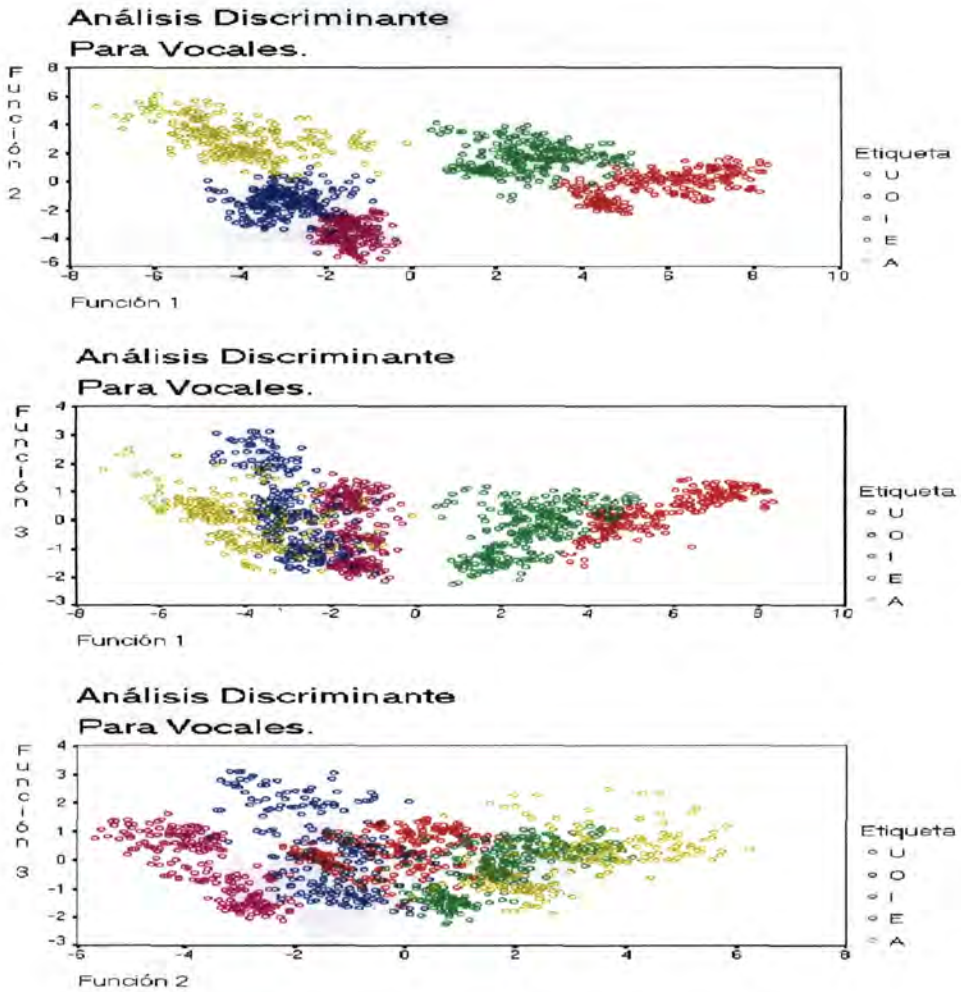


Figura 5.60

Resultados gráficos del análisis discriminante para vocales.

En la figura 5.61 se presentan los resultados del análisis discriminante para la clasificación de hablantes. Como se puede observar, la separabilidad es mucho peor que en el caso de las vocales. A pesar de que no forma parte de los objetivos de este trabajo, resulta interesante constatar la mejora que se produce respecto a los gráficos del apartado anterior, que aunque no empleábamos únicamente con esta finalidad, si estaba contemplada.

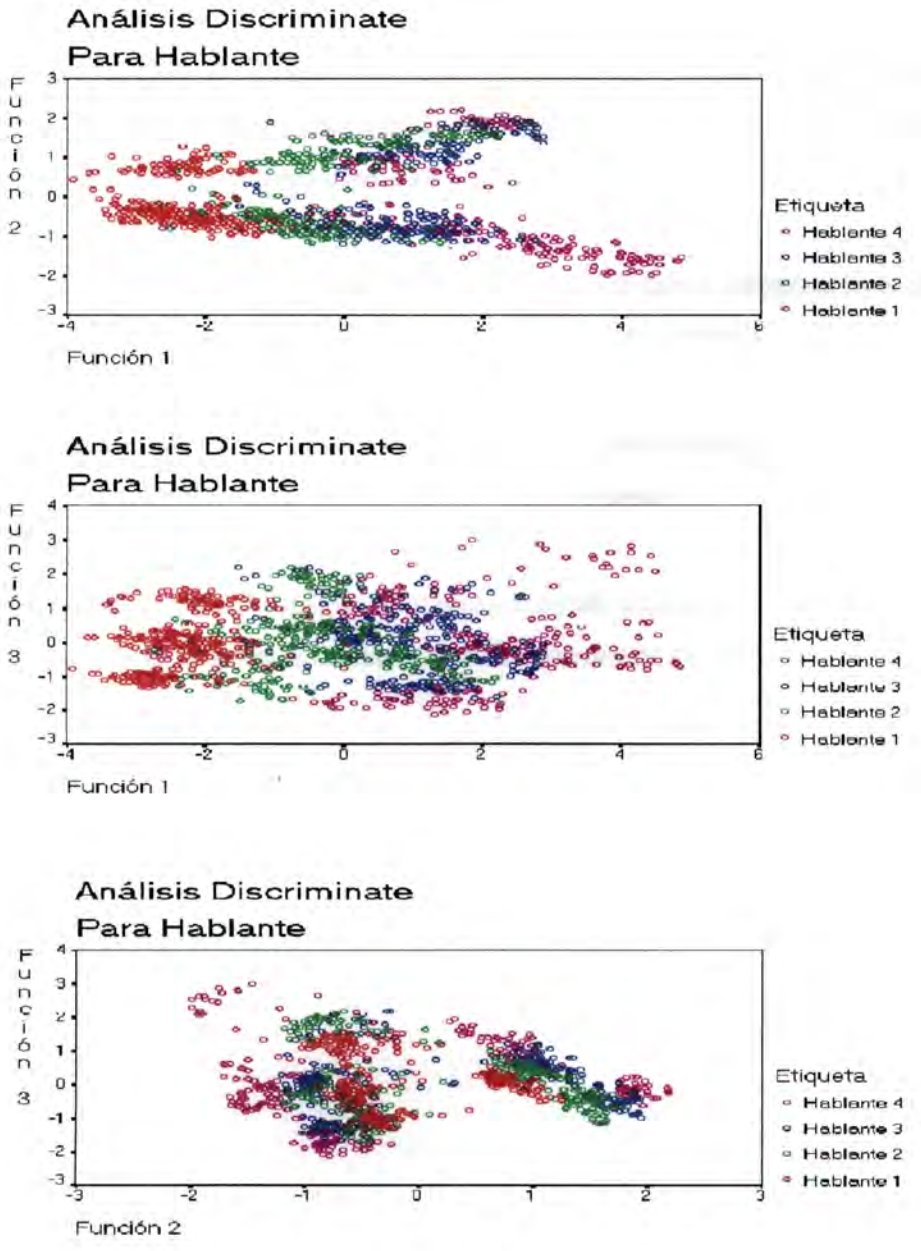


Figura 5.61

Resultados gráficos del análisis discriminante para hablantes.

5.5.5 CONCLUSIONES

Aplicando la metodología propuesta, se puede realizar un estudio de las posiciones de los formantes en varios hablantes.

Como es sabido, el punto de articulación de las consonantes adyacentes a las vocales, determina las diferentes posiciones que pueden adoptar los formantes.

El tono de voz de cada persona influye directamente en la altura de los formantes que presentan los espectros, esto hace necesario una fase de normalización para conseguir una correcta caracterización de los parámetros espectrales de mayor interés.

Es posible realizar una adecuada distinción vocálica acudiendo a comparaciones en el plano F1-F2 y apoyándose en caso de necesidad en los planos F1-F3 y F2-F3.

La metodología propuesta se basa en clasificaciones de datos estáticos de formantes de voz, pero ayuda a la comprensión de la evolución de estos formantes.

La utilización de técnicas de análisis discriminante ayuda a conseguir una buena clasificación vocálica, que mejora la obtenida mediante la utilización directa de formantes.

5. 6 SONIDOS NO VOCÁLICOS EN VARIOS HABLANTES

5.6.2 INTRODUCCIÓN

Aunque las consonantes en sí mismas presentan características importantes que contribuyen a su identificación, resulta fundamental determinar las transiciones vocálicas que las rodean. En este apartado se realiza un estudio de la evolución de los formantes en las vocales coarticuladas con algunos sonidos consonánticos del castellano.

El trabajo aquí mostrado ha sido realizado con cinco hablantes de diferentes sexos y edades, utilizando la siguiente metodología:

1.- Realizar grabaciones de secuencias /VCV/ tales como 'apa', 'epe', 'ipi', 'opo', 'upu', 'aβa', 'eβe', etc.
2.- Obtener los espectros básicos y mejorados utilizando los métodos y algoritmos desarrollados en el capítulo anterior.
3.- Repetir el proceso con las grabaciones en las que no aparece una evolución de formantes bien definida.
4.- Generalizar y resumir las evoluciones de los formantes por cada sonido consonántico escogido.

En el apéndice se incluyen parte de los espectros más representativos obtenidos utilizando el proceso detallado. Los espectros han sido clasificados atendiendo al modo de articulación de los sonidos consonánticos. En cada caso se detalla la cuantía de las subidas o bajadas de las transiciones vocálicas medidas en Hercios. Puesto que el tono de la voz varía según el sexo, en cada caso se indica como H/M (Hombre/Mujer) esta característica, que influye en la altura frecuencial de los formantes.

A continuación se presentarán gráficamente los resúmenes de las evoluciones de los formantes hallados. En primer lugar atendiendo al punto de articulación (debido a la coincidencia en el locus que presentan los sonidos con similares puntos de articulación) en los grupos oclusivo sordo, fricativo sonoro y nasal; después se mostrarán los grupos laterales/vibrantes (líquidas) y fricativos/africado.

Por cada sonido consonántico se detallan las evoluciones de los tres primeros formantes en las vocales castellanas. En algunos casos se dibujan dos posibles evoluciones, puesto que así se han descubierto en el análisis de los espectros obtenidos.

Es importante recordar que diversas publicaciones [MAS75], [REP78] señalan la importancia de la vocal posterior a la consonante como elemento de identificación consonántica, por lo que en los resultados que a continuación se presentan conviene fijar una mayor atención en la vocal posterior que en la anterior en los grupos /VCV/.

En [QUIL93] se realiza un breve repaso de las evoluciones más significativas de los formantes ante algunos sonidos del español. La imprecisión y falta de completitud de los resultados conseguidos hasta el momento, nos dan una idea de la complejidad que presenta la investigación en este aspecto de la fonética acústica.

5.6.2 CONSONANTES BILABIALES

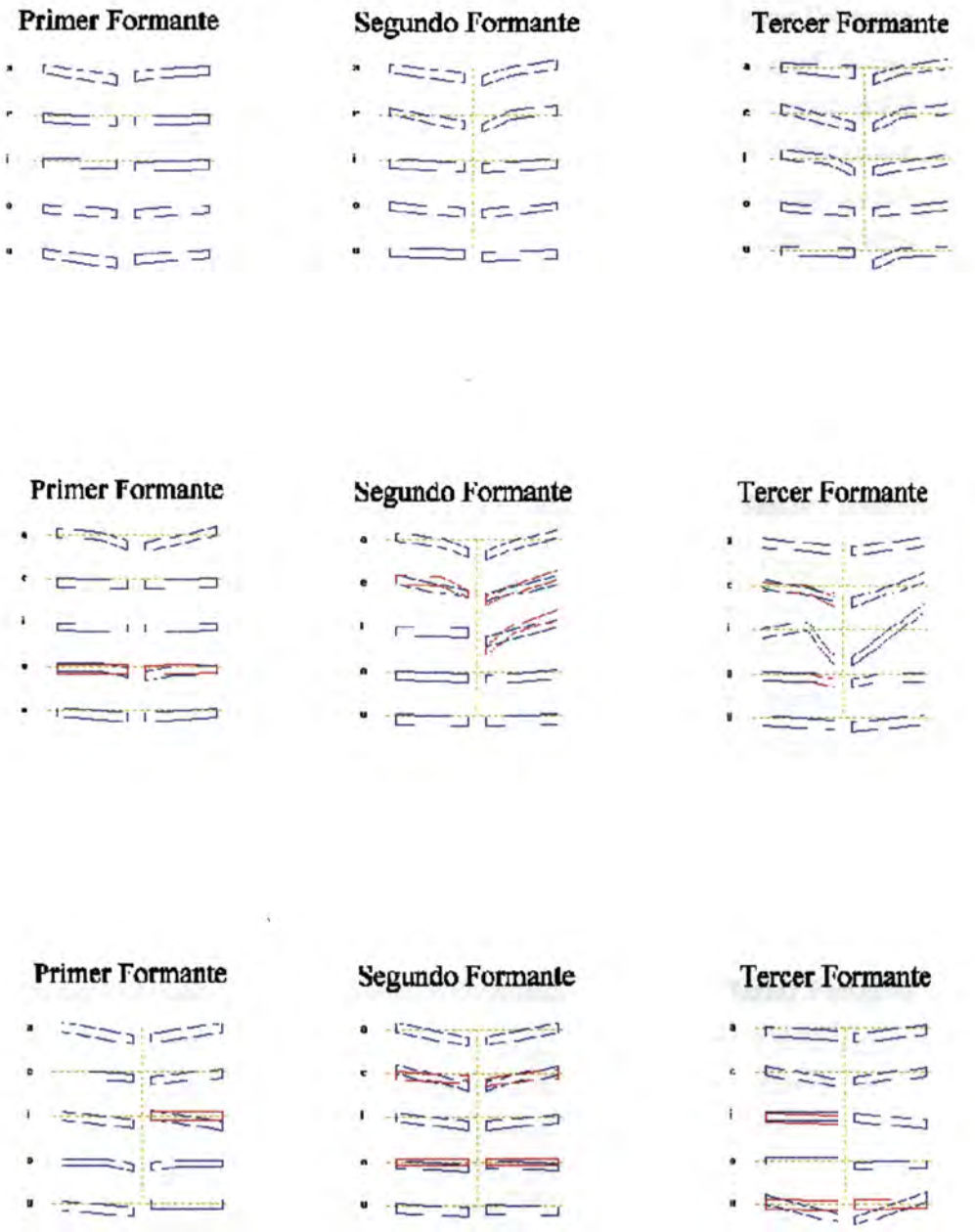


Figura 5.62

Resultados obtenidos de la evolución de los formantes alrededor de los sonidos bilabiales:

[p] imagen superior **[β] imagen central** **[m] imagen inferior**

5.6.3 CONSONANTES DENTALES/INTERDENTALES/ALVEOLARES

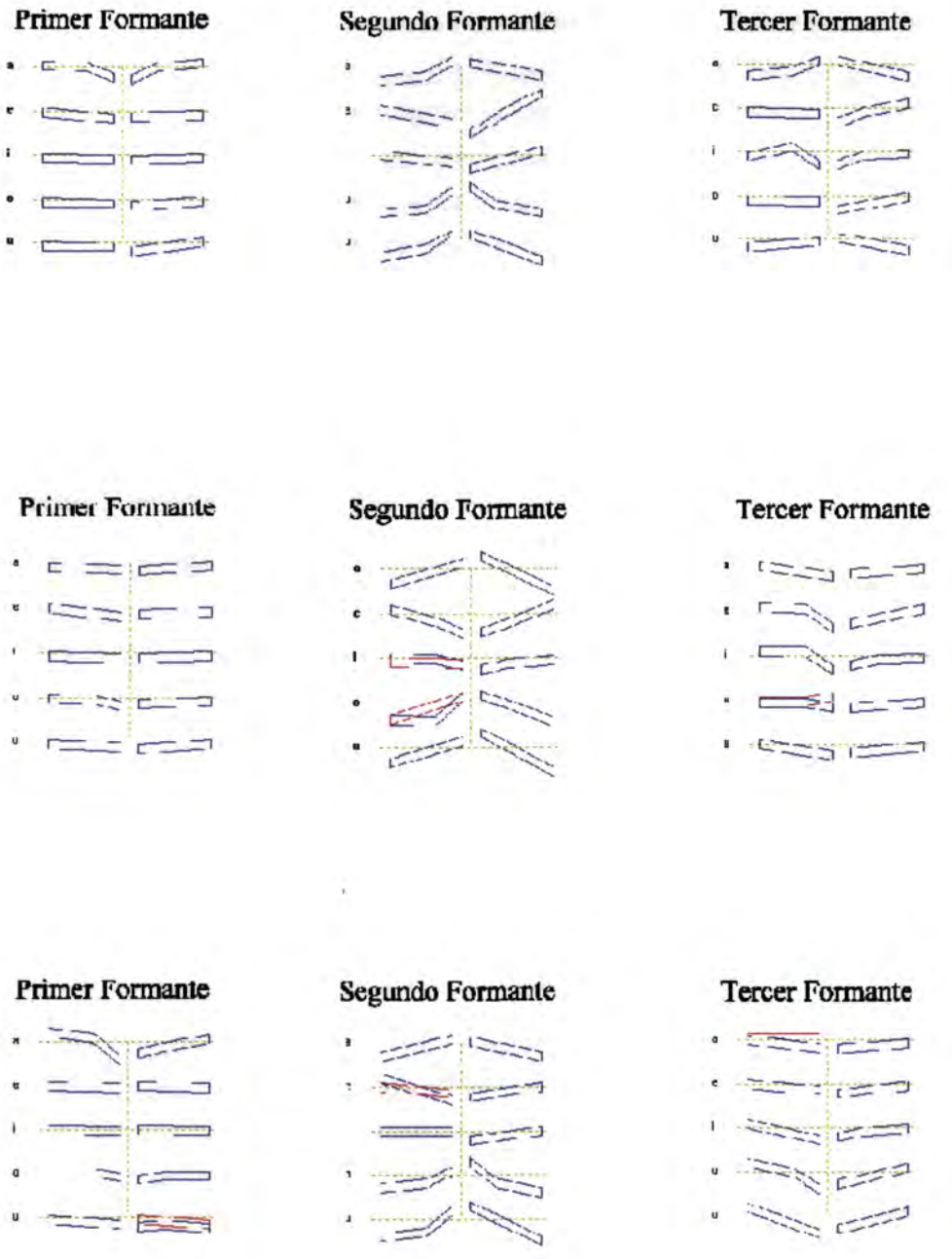


Figura 5.63

Resultados obtenidos de la evolución de los formantes alrededor de los sonidos:

[t] imagen superior [d.] imagen central [n] imagen inferior

5.6.4 CONSONANTES VELARES

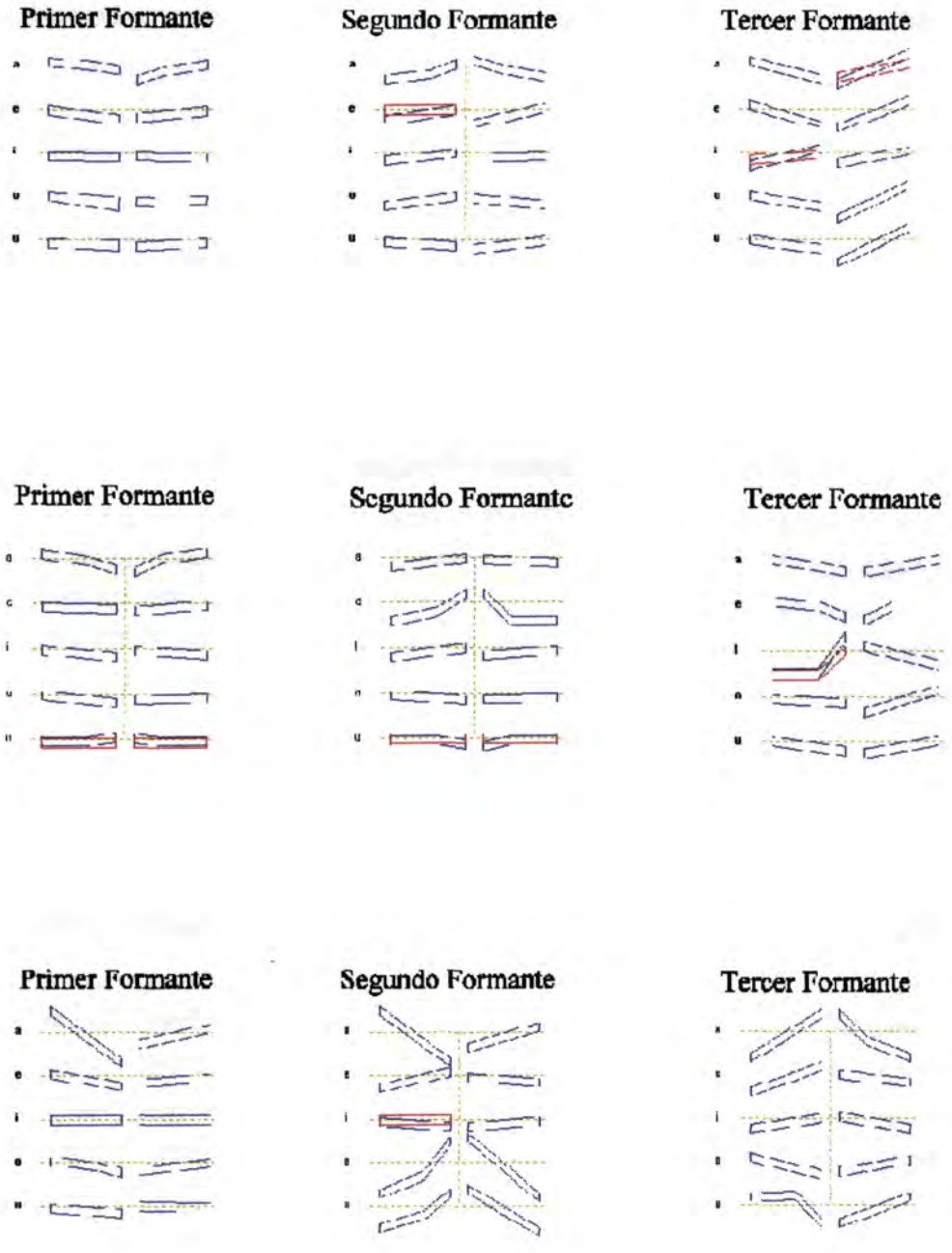


Figura 5.64

Resultados obtenidos de la evolución de los formantes alrededor de los sonidos velares:

[k] imagen superior [γ] imagen central [ɳ] imagen inferior

5.6.5 CONSONANTES FRICATIVAS/AFRICADA

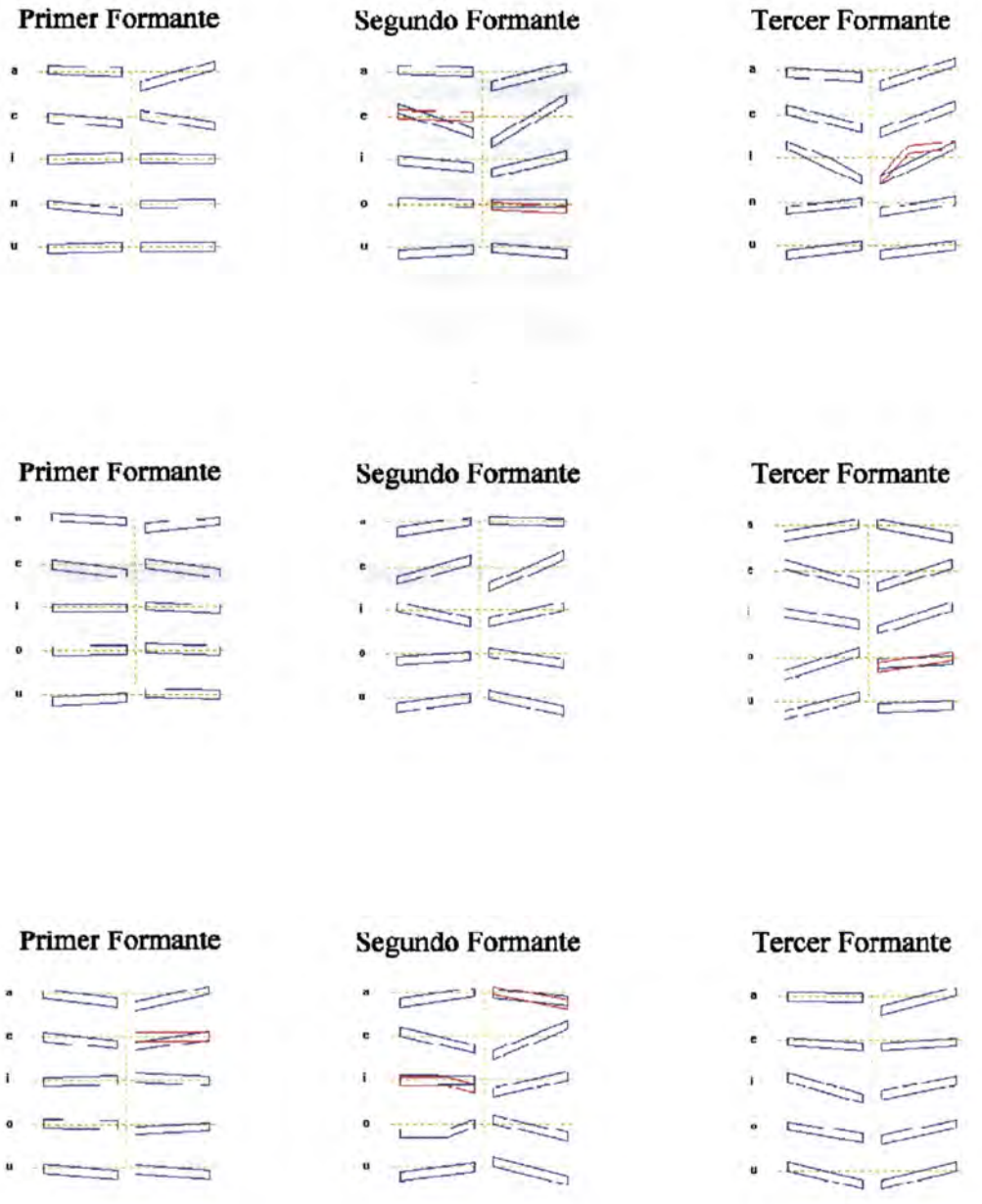


Figura 5.65

Resultados obtenidos de la evolución de los formantes alrededor de los sonidos fricativos:

[f] imagen superior [θ] imagen central [s] imagen inferior

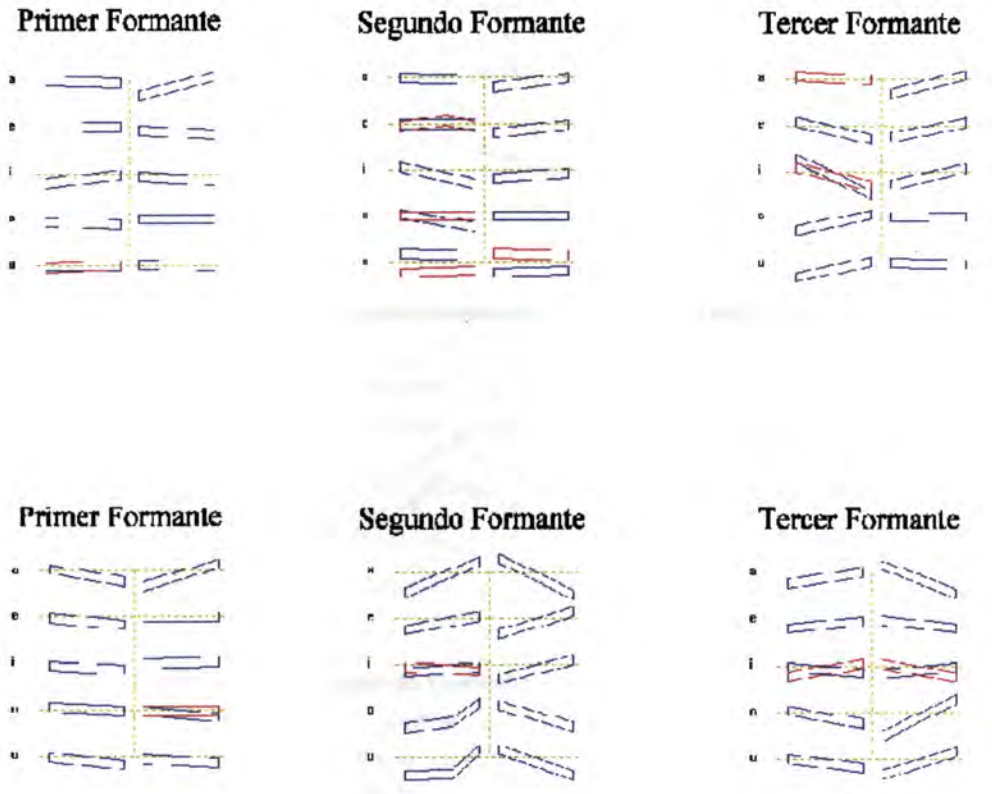


Figura 5.66

Resultados obtenidos de la evolución de los formantes alrededor de los sonidos fricativos/africado:

[x] imagen superior **[tʃ] imagen inferior**

5.6.6 CONSONANTES LÍQUIDAS

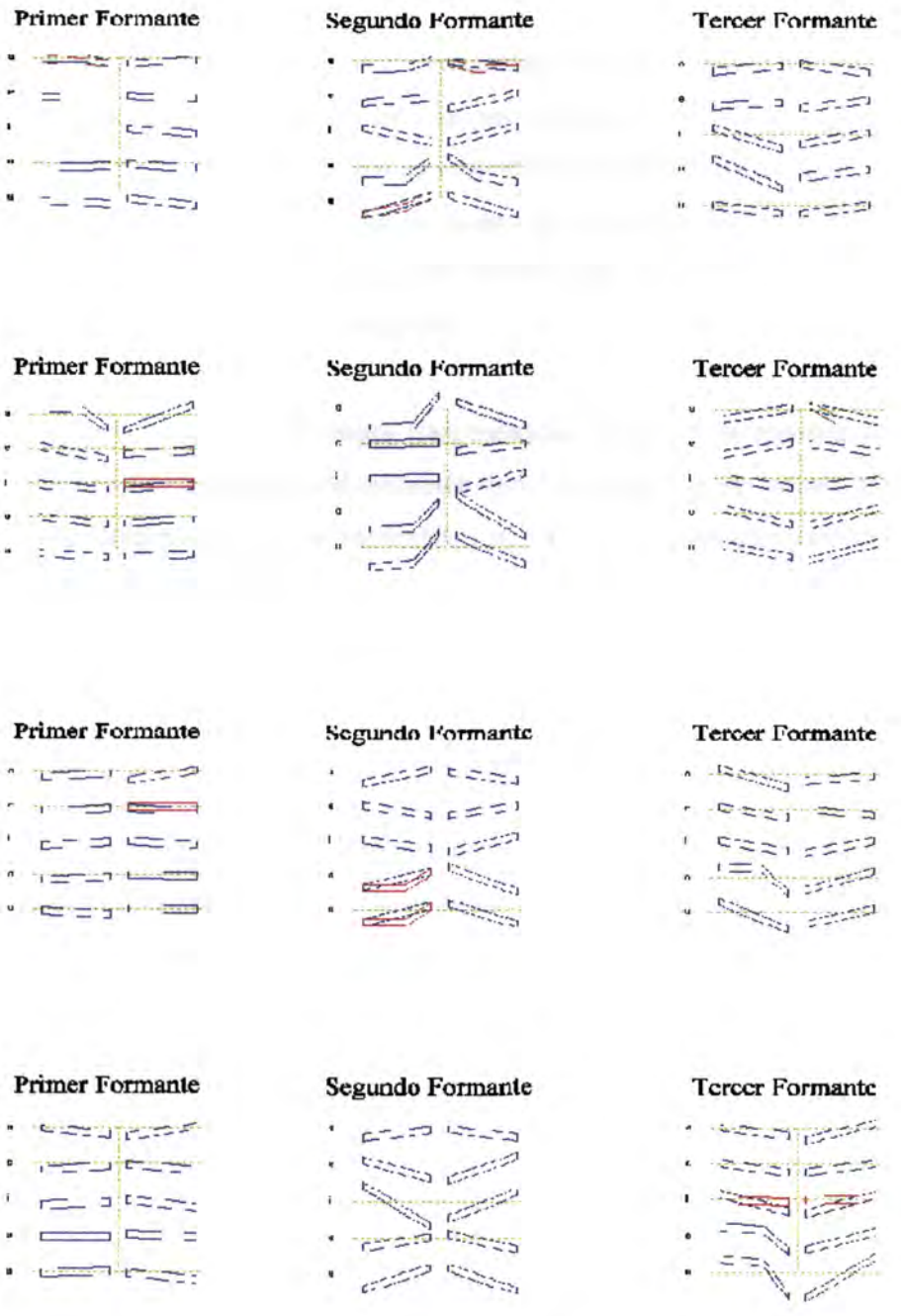


Figura 5.67

Resultados obtenidos de la evolución de los formantes alrededor de los sonidos líquidos:

[l] imagen superior **[λ] imagen segunda**

[r] imagen tercera **[rr] imagen inferior**

5.6.7 CONCLUSIONES

Existen dos conclusiones fundamentales que se obtienen tras la realización de este trabajo:

- La primera de ellas se refiere a la gran cantidad de esfuerzo que supone la realización de grabaciones con varios hablantes, su procesamiento informático y posterior análisis. Concretamente, el desarrollo que se presenta en este apartado ha requerido de unas 100 horas de dedicación (sin ponderar el tiempo de creación de las herramientas informáticas utilizadas). Esto supone un importante esfuerzo que no se refleja convenientemente en la importancia de los resultados obtenidos.

- La segunda conclusión, de mayor trascendencia, se basa en la constatación de que existe una enorme variabilidad en la evolución de los formantes de las vocales coarticuladas con sonidos consonánticos, esta variabilidad nos impide ofrecer unos resultados generales y fiables; de hecho, los resultados de este trabajo se pueden tomar como una pauta, pero no como una regla de comportamiento de los formantes de las vocales junto a sonidos consonánticos. Un estudio profundo en esta dirección daría cabida sin lugar a dudas a una tesis en el campo de la fonética acústica del idioma analizado.

Aunque las relaciones entre el punto de articulación y la posición del locus teórico se mantienen, los resultados obtenidos contradicen la posibilidad que en [MAR94] se presenta de fijar el locus en posiciones fijas y concretas.

5.7 CONCLUSIONES

En el trabajo realizado sobre la situación y evolución de formantes, se obtiene como primera y más importante conclusión la constatación del hecho de que existen grandes diferencias entre el estudio de las posiciones estáticas de los formantes y el de sus evoluciones. Mientras que la posición de los formantes se puede determinar con relativa facilidad, la obtención de sus evoluciones a lo largo del tiempo resulta más costosa y menos fiable de lo que cabría esperar.

La utilización de varios hablantes en las grabaciones genera resultados más dispares, produciéndose mayores dispersiones en las clasificaciones y generalizaciones obtenidas, especialmente entre individuos de diferentes sexos.

La representación visual de los formantes en planos de proyección (F1-F2, F1-F3, F2-F3), nos ofrece una visión complementaria a los espectros que ayuda a clasificar e identificar los sonidos vocálicos. Con la utilización de técnicas de análisis discriminante se consigue mejorar las clasificaciones obtenidas a partir de la posición de los formantes.

Las evoluciones de los formantes, aunque siguen unas reglas y pautas generales, presentan mucha variación entre distintas realizaciones del habla. Para poder realizar análisis representativos es necesario obtener, confrontar y comparar un gran número de grabaciones de voz de diferentes hablantes pronunciando sonidos básicos rodeados de diferentes contextos.

En esta tesis se ofrece abundante y variado material en forma de espectros típicos (situados en el apéndice), generalizaciones de la evolución de los formantes, planos bidimensionales de situación de vocales, etc. Estos datos y resultados, junto a la metodología y herramientas empleados, pueden servir de base para enfocar un trabajo de objetivos más ambiciosos en el campo de la fonética.

Las cualidades de los métodos y algoritmos desarrollados en el capítulo anterior han sido validadas en los diferentes estudios aquí realizados, pudiéndose afirmar que se proporciona un buen soporte para el análisis y estudio de diversas características espectrales del habla, especialmente en la detección de los formantes.

6 CONCLUSIONES

6.1 CONCLUSIONES GENERALES

La principal conclusión que se obtiene del trabajo realizado es la confirmación de nuestra hipótesis inicial, es decir, partiendo del método de predicción lineal, han sido ideados algoritmos de tratamiento de señal que permiten obtener una buena estimación de características espectrales significativas de la voz, especialmente en la detección de los formantes que se producen en el habla.

Los estudios realizados en el campo de la fonética acústica han llevado a alcanzar dos metas principales: por una parte han servido para validar los métodos y algoritmos de tratamiento de señal ideados y desarrollados a lo largo de la tesis. Por otro lado, han cubierto sobradamente uno de los objetivos iniciales de este trabajo, consistente en aportar metodologías y resultados que sirvan de base para futuras investigaciones en el campo de la fonética.

Entre los objetivos iniciales de la tesis se encontraba el compromiso de enfocar las investigaciones hacia la extracción de formantes, debido a la gran importancia que suponen los avances en esta materia. Se puede afirmar que el mayor peso de los desarrollos realizados se centra en la creación de métodos y algoritmos de detección de formantes y su posterior validación en el ámbito de la fonética castellana.

La decisión de trabajar con funciones espectrales suavizadas ha resultado muy adecuada para la estimación de los formantes de voz. Partiendo de estas funciones espectrales se han ideado diferentes etapas que van detectando y resaltando los formantes del habla haciendo uso de transformaciones no lineales basadas en métodos algorítmicos.

Para la caracterización y representación de sonidos sonoros se han ideado dos funciones espectrales de gran interés, una de ellas debido a su idoneidad en la extracción automática de formantes, y la otra, por sus especiales características para la representación visual de los espectros del habla.

Con los métodos ideados para la extracción de formantes, se consiguen buenos resultados sin utilizar mecanismos de suavizado de señal. Debido a que los algoritmos empleados no hacen uso de la información proporcionada por la evolución de las funciones espectrales a lo largo del tiempo, queda abierta la posibilidad de incluir etapas que mejoren los resultados por esta vía.

La elección de una escala de colores adecuada es muy importante para conseguir una correcta visualización de los espectros de voz. Esto es debido a que de esta manera se facilita la identificación de las zonas más representativas del gráfico obtenido.

Un objetivo primordial para el desarrollo de los distintos campos del tratamiento automático del habla, es la obtención de métodos fiables de caracterización de los sonidos básicos de un idioma. Para la consecución de este objetivo, se requiere la aportación de nuevas ideas y enfoques en el área del tratamiento de la señal, particularizando las investigaciones en las características propias de la voz.

Los algoritmos desarrollados proporcionan en su conjunto una buena calidad en la visualización de los espectros de voz, calidad basada en la correcta determinación de las características espectrales buscadas, sin embargo, los tiempos de respuesta no son lo suficientemente cortos y predecibles como para ser utilizados como núcleo de aplicaciones de voz con fuertes limitaciones temporales.

Delimitar con exactitud las diferentes zonas de un espectro de voz resulta muy complicado, debido a la enorme gama de variaciones con las que nos podemos encontrar en el habla conexa. Esta dificultad se nos ha presentado especialmente en la determinación de las características de sordez y sonoridad de los sonidos.

La representación visual de los formantes en planos de proyección (F1-F2, F1-F3, F2-F3), nos ofrece una visión complementaria a los espectros, que ayuda a clasificar e identificar sonidos vocálicos. Los mapas bidimensionales nos muestran la gran variabilidad existente entre las distintas realizaciones del habla producidas por diferentes hablantes. Esta variabilidad se

manifiesta especialmente en las distintas alturas frecuenciales que presentan los formantes, debido a las diversas entonaciones que se producen en la voz. Con la utilización de técnicas de análisis discriminante se consigue mejorar las clasificaciones obtenidas a partir de las posiciones de los formantes.

Las visualizaciones de los sonidos en mapas bidimensionales o tridimensionales pueden ser mejoradas normalizando los datos de cada hablante con su frecuencia fundamental (F0). Es posible realizar una adecuada distinción vocálica acudiendo al plano F1-F2 y apoyándose en caso de necesidad en los planos F1-F3 y F2-F3. Además, estas representaciones correctamente diseñadas e interpretadas nos ayudan a comprender los efectos de coarticulación que se producen en el habla.

Para poder interpretar los espectros de voz, es necesario comprender la evolución de los formantes en las vocales adyacentes a los sonidos consonánticos. En el trabajo realizado sobre la situación y evolución de formantes, se obtiene como conclusión fundamental la constatación del hecho de que existen grandes diferencias entre el estudio de las posiciones estáticas de los formantes y el de sus evoluciones. Mientras que la posición de los formantes se puede determinar con relativa facilidad, la obtención de sus evoluciones a lo largo del tiempo resulta más costosa y menos fiable de lo que cabría esperar.

Resulta muy adecuado investigar empíricamente el campo del tratamiento de la voz, para lo cual es necesario utilizar potentes herramientas informáticas. En nuestro caso, estas herramientas han sido desarrolladas en su totalidad según iban apareciendo nuevas ideas o necesidades. La experiencia nos ha mostrado como elección más adecuada el uso de entornos visuales orientados a objetos, utilizados para realizar prototipos basados en el modelo en espiral de ingeniería del software.

6.2 APORTACIONES ORIGINALES

En este apartado se resumen las aportaciones de la tesis que resultan novedosas frente a los trabajos realizados en el área hasta el momento. Las aportaciones originales se basan en los resultados más interesantes conseguidos mediante los diversos estudios experimentales llevados a cabo.

En primer lugar, la obtención de los formantes del habla se ha basado en el método de predicción lineal, haciéndose una búsqueda de polos fuera de la zona habitual (el círculo unidad). Esta decisión ha facilitado la determinación de formantes al poderse trabajar con funciones espectrales suavizadas.

Además de la creación de un método propio de obtención de formantes, se ha desarrollado un algoritmo que nos permite conseguir una función especialmente diseñada para ser utilizada en la visualización de espectros del habla.

Los mapas tridimensionales de sonidos vocálicos que se aportan en la tesis, sirven de modelo para la extensión de las frecuentes clasificaciones bidimensionales que se utilizan en las publicaciones especializadas de fonética acústica. El empleo de una tercera dimensión permite complementar la información tradicional usada en las representaciones vocálicas.

Aunque el estudio realizado sobre la evolución de los formantes en situaciones de coarticulación no es completo, sí que se puede considerar como referencia innovadora para el desarrollo de trabajos más elaborados que se basen en los métodos y herramientas originales empleados en la tesis.

La aplicación informática, producida en paralelo al desarrollo de los estudios y trabajos de investigación, puede ser usada por cualquier persona que se interese en las áreas de tratamiento de señal y fonética acústica. Sus principales ventajas son las siguientes:

- Funciona sin necesidad de utilizar hardware específico.
- Se proporciona como una aplicación abierta, donde es posible modificar o ampliar cualquier funcionalidad que se considere adecuada.
- Está realizada utilizando técnicas de programación visual y orientada a objetos, aportándose módulos reutilizables por otros entornos.

6.3. AMPLIACIONES PROPUESTAS

Las posibilidades de ampliación y continuación de los trabajos desarrollados a lo largo de esta tesis se dividen en dos áreas fundamentales: en el campo del tratamiento de señal, resultaría interesante incorporar un sistema de separación de polos muy cercanos que, con los algoritmos desarrollados, en ocasiones no se identifican como formantes diferentes. En la parte correspondiente a los estudios fonéticos, podría hacerse uso de los algoritmos, herramientas, métodos, datos y resultados proporcionados para iniciar investigaciones más amplias y ambiciosas en el campo de la fonética acústica.

En especial resultaría muy adecuado extender las investigaciones y desarrollos en los siguientes campos:

- Determinación del tono fundamental de la voz como base de normalización en la representación de mapas tridimensionales de sonidos.

- Obtención de mapas tridimensionales de sonidos normalizados, partiendo de datos obtenidos utilizando un conjunto amplio y representativo de hablantes.

- Diseño de métodos robustos de suavizado de señal, con el fin de determinar más apropiadamente las trayectorias de los formantes del habla.

- Creación de algoritmos que permitan realizar una adecuada separación de los sonidos sonoros, los sordos y los silencios en el habla conexa.

- Análisis de las características acústicas de los sonidos sordos de forma aislada, es decir, sin acudir a la información que proporciona la evolución de los formantes de los sonidos adyacentes.

- Ampliación de los estudios realizados sobre coarticulación, empleando un número superior de hablantes y analizando una mayor variedad de sonidos, con el fin de conseguir resultados más completos y fiables.

- Utilización de los resultados obtenidos para desarrollar un prototipo de reconocedor de voz basado en el análisis de secuencias cortas de sonidos

7 BIBLIOGRAFÍA

[ASS95a] P.F. Assmann, W.F. Katz, K.M. Jenouri, P.W Hamilton, "Identification of natural and synthetised vowels produced by children and adults: Effects of formant frequency variation", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 4pSC3, 1995

[ASS95b] P.F. Assmann, "The role of formant transitions in the perception of concurrent vowels", *Journal of the Acoustic Society of America*, Vol. 97 (1), Enero 1995, pp. 575-584

[ASS96] P.F. Assmann, "Modeling the perception of concurrent vowels: Role of formant transitions", *Journal of the Acoustic Society of America*, Vol. 100 (2), Agosto 1996, pp. 1141-1152

[ATA71] B.S. Atal, S.L Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", *Journal of the Acoustic Society of America*, Vol. 50, 1971, pp. 637-655

[BLA82] R.A. Bladon, "Arguments against formants in the auditory representation of speech", *The Representation of Speech in the Peripheral Auditory System*, 1982, pp. 95-102

[BLU79] S.E. Blumstein, K.N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *Journal of the Acoustic Society of America*, Vol. 66, 1979, pp. 1001-1017

[BLU80] S.E. Blumstein, K.N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments", *Journal of the Acoustic Society of America*, Vol. 67 , 1980, pp. 648-662

[BON96] A. Bonneau, L. Djezzar, Y. Laprie, "Perception of the place of articulation of French stop bursts", *Journal of the Acoustic Society of America*, Vol. 100 (1), Julio 1996, pp. 555-564

[BOO96] A. Boothroyd, B. Mulhearn, J. Gong, J. Ostroff, "Effects of spectral smearing on phoneme and word recognition", *Journal of the Acoustic Society of America*, Vol.100 (3), Septiembre 1996, pp. 1807-1818

[BRO89] D.J. Broad, F. Clermont, "Formant estimation by linear transformation of the LPC cepstrum", *Journal of the Acoustic Society of America*, Vol. 86, 1995, pp. 2013-2017

[BUS95] P.A. Busby, G.L Plant, "Formant frequency values of vowels produced by preadolescent boys and girls", *Journal of the Acoustic Society of America*, Vol. 97(4), 1995, pp. 2603-2607

[CAN74] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 22(2), 1974, pp. 135-141

[CAS87] F. Casacubierta, E. Vidal, *Reconocimiento automático del habla*, Marcombo, 1987

[CAS90] F. Casacubierta, E. Vidal, *Reconocimiento automático del habla*, Estudios de Fonética Experimental, 1990, Vol. 4, pp. 167-178

[COA93] P. Coad, J. Nicola, *Object-oriented programming*, Yourdon Press, 1993

[COC67] W.T. Cochran, *What is the Fast Fourier Transform?*, IEEE Trans. Audio Electroacoust., Vol. AU-15, 1967, pp. 45-55

[COO93] M. Cooke, S. Beet, Crawford M., *Visual Representations of speech signals*, John Wiley, 1993

[COU95] M.P. Coughlin, D. Kewley-Pot, L.E. Humes, "The relation between identification and discrimination of vowels by young normal-hearing and elderly hearing-impaired listeners", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 4pSC1, 1995

[CHA96] C. Calvert, *Delphi2 unleashed*, Borland Press, 1996

[CHE95] M.Y. Chen, "Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers", *Journal of the Acoustic Society of America*, Vol. 98 (5), Noviembre 1995, pp. 2443-2453

[FER93] A.M. Fernández, "Estudio del campo de dispersión de las vocales castellanas", *Estudios de Fonética Experimental*, Vol. 5, 1993, pp. 129-162

[FOL92] G.B. Folland, *Fourier analysis and its applications*, Wadsworth & Brook, 1992

[FRE93] J. Freeman, D. Skapura, *Redes neuronales, algoritmos, aplicaciones y técnicas de programación*, Addison-Wesley/Díaz de Santos, 1993

[GOT80] T.L. Gottfried, W. Strange, "Identification of coarticulated vowels", *Journal of the Acoustic Society of America*, Vol. 68, 1980, pp. 1626-1635

[GRA91] B. Grady, *Object oriented design with applications*, The Benjamin/Cumming, 1991

[HEI93] R. Heimlich, *Sound Blaster: the official book*, McGraw Hill, 1993

[HEU96] H.V. Heuvel, B.Cranen, T. Rietveld, "Speaker variability in the coarticulation of /a,i,u/", *Speech Communication*, Vol. 18, 1996, pp. 113-130

[HIL95] J. Hillenbrand, L.A. Getty, M.J. Clark, K. Wheeler, "Acoustic characteristics of American English vowels", *Journal of the Acoustic Society of America*, Vol. 97 (5), Mayo 1995, pp. 3099-3111

[HIL95b] J.R. Hilera, V. Martínez, *Redes neuronales artificiales, fundamentos, modelos y aplicaciones*, Ra-Ma, 1995

[HOR85] P. Horman, D. King, *Expert systems*, John Wiley, 1985

[JAC90] P. Jackson, *Introduction to expert systems*, Addison-Wesley, 1990

[KAT95] W.F. Katz, P.F. Assmann, K.M. Jenouri, "Identification of natural and synthetized vowels produced by children and adults: Effects of fundamental frequency variation", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 4pSC2, 1995

[KEW93] W. Kewin, *El libro de la Sound Blaster*, ANAYA, 1993

[KEW95] D. Kewley-Port, "Thresholds for formant-frequency discrimination of vowels in consonantal context", *The Journal of the Acoustic Society of America*, Vol. 97 (5), Mayo 1995, pp. 3139-3146

[KOE46] W. Koening, H.K. Dunn, L.Y. Lacy, "The sound spectrograph", *The Journal of the Acoustic Society of America*, Vol. 18(1), 1946, pp. 19-28

[KUS95] R.E. Kushner, "Analysis and perception of voice similarities among family members", *130th Meeting: Acoustic Society of America, Speech Communication: Studies of Voice,* 3pSC1, 1995

[LEE96] M.R. Leek, V. Summers, "Reduced frequency selectivity and the preservation of spectral contrast in noise", *Journal of the Acoustic Society of America*, Vol. 100 (3), Septiembre 1996, pp. 1796-1806

[LIN67] B. Lindblom, M. Studdert-Kennedy, "On the role of formant transitions in vowel recognition", *Journal of the Acoustic Society of America*, Vol. 42, 1967, pp. 830-843

[MAK75] J. Makhoul, "Linear prediction: a tutorial review", *Proceedings of the IEEE*, Vol. 63(4), 1975, pp. 561-580

[MAK85] J. Makhoul, S. Roucos, H. Gish, "Vector Quantization in speech coding", *Proceedings of the IEEE*, Vol. 73, 1985, pp. 1551-1588

[MAR90a] E. Martínez, "Una utilidad fonética: la carta de formantes por ordenador", *Estudios de Fonética Experimental*, Vol. 4, 1990, pp. 179-193

[MAR90b] R. Marti, J., "Situación actual de la síntesis de voz", *Estudios de Fonética Experimental*, 1990, Vol. 4, pp. 167-178

[MAR94] E. Martínez Celdrán, *Fonética*, Martínez Celdrán E., Teide, 1994

[MAS75] D.W. Massaro, "Preperceptual images, processing time, and perceptual units in speech perception", *Understanding language (Academic Press New York)*, 1975, pp. 125-150

[MED85] C. Medina Casado, "El espectrógrafo. Su utilización para la adquisición de una nueva lengua", *CAUCE*, Nº 8, 1985, pp. 217-227

[MEM78] P. Mermelstein, "Difference limens for formant frequencies on steady-state and consonant-bound formants", *Journal of the Acoustic Society of America*, Vol. 63, 1978, pp. 572-580

[MON83] R.B. Monsen, "General effects of deafness on phonation and articulation, Speech of the Hearing Impaired", *(University Park, Baltimore)*, pp. 23-24

[MOR90] Mª A. Moreno, "Transiciones vocálicas y punto de articulación consonántico", *Estudios de Fonética Experimental*, Vol. 4, 1990, pp. 50-102

[MUJ90] E. Mújica, Mª M. Santos, J. Herraiz, "Duración de las transiciones en las oclusivas sordas del castellano", *Estudios de Fonética Experimental*, Vol. 4, 1990, pp. 103-122

[OHD95] R.N. Ohde, K.L. Haley, H.K. Vorperian, "A developmental study of the perception of onset spectra for stop consonants in different vowel environments", *Journal of the Acoustic Society of America*, Vol. 97 (6), Junio 1995, pp. 3800-3812

[PAR86] T. Parsons, *Voice and speech processing*, Mc Graw Hill, 1986

[PET52] G.E. Peterson, H.L. Barney, "Control methods used in a study of the vowels", *Journal of the Acoustic Society of America*, Vol. 24, 1952, pp. 175-184

[PIC95] J.M. Pickett, H. Bunell, S. Revoile, "Phonetics of intervocalic consonant perception: retrospect and prospect", *Phonetica*, Vol. 52, 1995, pp. 1-40

[PLA95] F. Plante, W.A. Ainsworth, "Formant tracking using reassigned spectrum", *EUROSPEECH 95*, 1995, pp. 741-744

[QUI93] A. Quilis, *Tratado de fonología y fonética españolas*, Gredos, 1993

[RAB70] L.R. Rabiner, R.W. Schafer, "System for automatic formant analysis of voiced speech", *Journal of the Acoustic Society of America*, Vol. 47(2), 1970, pp. 634-648

[RAB78] L.R. Rabiner, R.W. Schafer, *Digital processing of speech signals*, Prentice Hall, 1978

[RAB89] L.R. Rabiner, "A tutorial on Hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol. 77, Nº 2, 1989, pp. 257-286

[RAB93] L.R. Rabiner, *Fundamentals of speech recognition*, Biing-Hwang Juang, Prentice Hall, 1993

[RAN95] M. Rangoussi, A. Delopoulos, "Recognition of unvoiced stops from their time-frequency representation", *International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, 1995, pp. 792-795

[REP78] B.H. Repp, "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants", *Perception Psychophysics*, Vol. 24, 1978, pp. 471-485

[ROM88] J. Romero, "Campos de dispersión auditivos de las vocales del castellano. Percepción de las vocales", *Estudios de Fonética Experimental*, Vol. 3, 1988, pp. 86-95

[ROS96] D. Rossiter, D.M. Howard, M. DeCosta, "Voice development under training with and without the influence of real-time visually presented biofeedback", *Journal of the Acoustic Society of America,* Vol. 99 (5), Mayo 1996, pp. 3253-3256

[ROW92] C. Rowden, *Speech processing*, Mc Graw Hill, 1992

[SAB84] M.J. Sabater, "La experimentación en fonética y fonología", *Estudios de Fonética Experimental*, 1984, Vol. 1, pp. 1-70

[SCH70] R.W. Schafer, L.R. Rabiner, "System for automatic formant analysis of voiced speech", *The Journal of the Acoustic Society of America*, Vol. 47, N° 2, 1970, pp. 634-648

[SCH95] P. Schmid., E. Barnard, Robust , "N-best formant tracking", *EUROSPEECH95*, 1995, pp. 737-740

[SLI95] J. Slifka, T.R. Anderson, "Speaker modification with LPC pole analysis", *International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, 1995, pp. 664-667

[SMI94] R. Smits, "Accuracy of quasistationary analysis of highly dynamic speech signals", *Journal of the Acoustic Society of America*, Vol. 96(6), 1994, pp. 3401-3415

[STR89] W. Strange, "Dynamic specification of coarticulated vowels spoken in sentence context", *Journal of the Acoustic Society of America*, Vol. 85, 1989, pp. 2135-2153

[SU74] L.S. Su, K.P. Li, K.S. Fu, "Identification of speakers by use of nasal coarticulation", *Journal of the Acoustic Society of America,* Vol. 56, 1974, pp. 1867-1882

[TOH92] Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisaka, *Speech perception, production and linguistic structure*, IOS Press, 1992

[TOK93] S. Tokuma, "Some arguments on vowel formant shift", *Speech, Hearing and Language: Work in Progress,* UCL, Vol. 7, 1993, pp. 233-254

[TRE95] S.A. Trent, "Voice quality: Listener identification of African-American versus Caucasian speakers", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 3pSC2, 1995

[WAN95] R. Wang, W.J. Strong, "Acoustic study of acted emotions in speech", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice,* 3pSC4, 1995

[WAT90] R. L. Watrous, "Current status of Peterson-Barney vowel formant data", *Journal of the Acoustic Society of America*, Vol. 89 (5), Mayo 1991, pp. 2459-2460

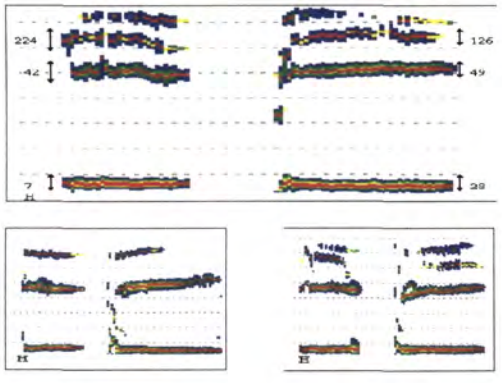
[NAG94] I. Nagayama, N. Akamatsu, T. Yoshimo, "Phonetic visualization for speech training system by using neural network", *ICSLP94,* Yokohama S32-22, pp. 2027-2030

[ZAH93] S.A. Zahorian, A.J. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels", *Journal of the Acoustic Society of America*, Vol. 94 (4), 1993, pp. 1966-1982

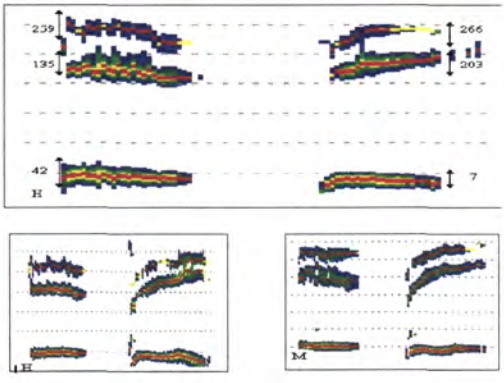
A APÉNDICE

[p]

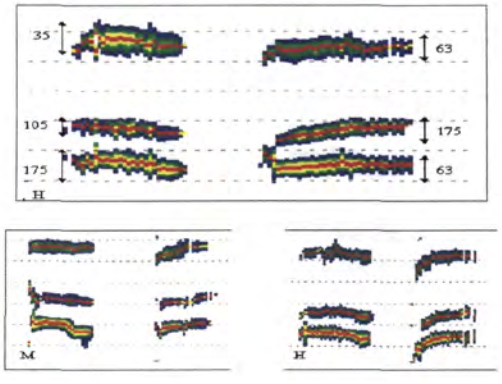
ipi



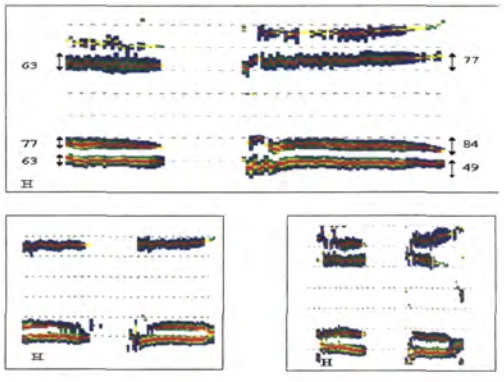
epe



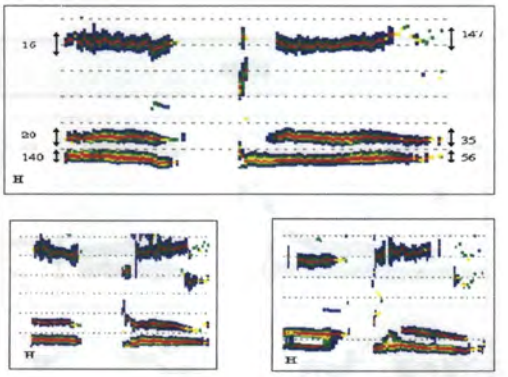
apa



opo

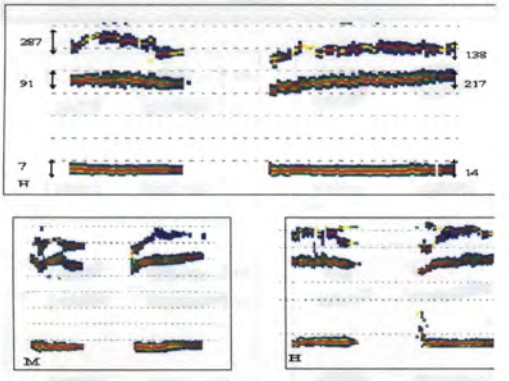


upu

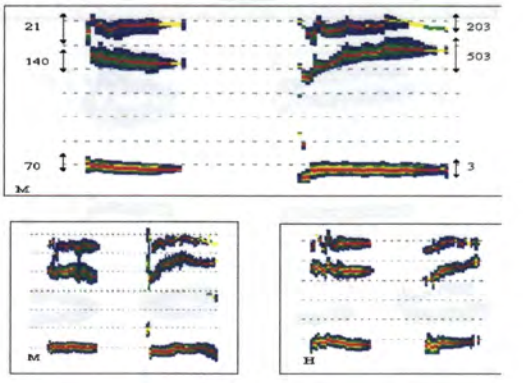


[t]

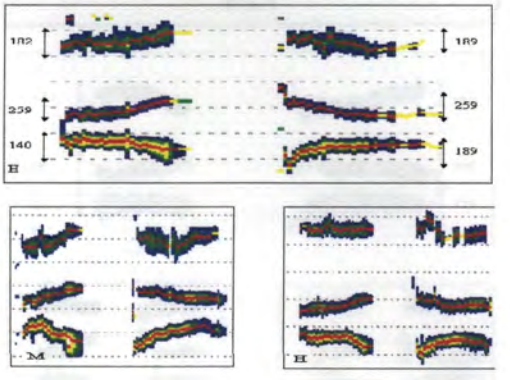
iti



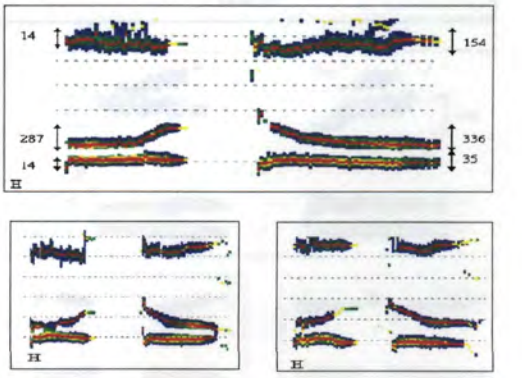
ete



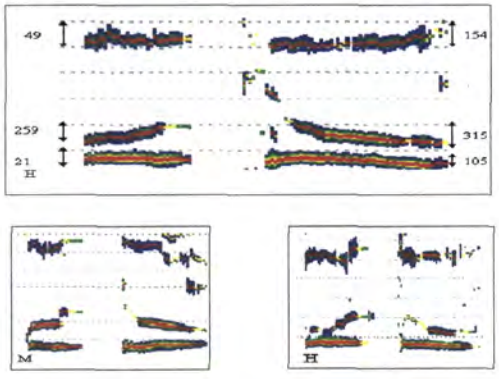
ata



oto

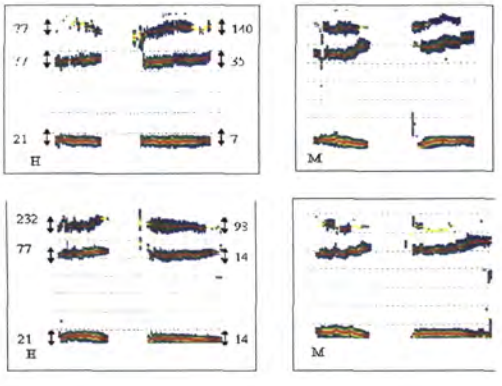


utu

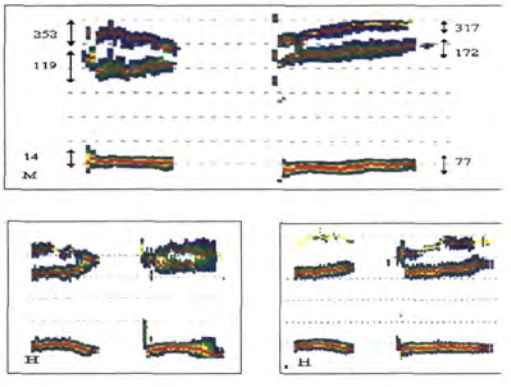


[k]

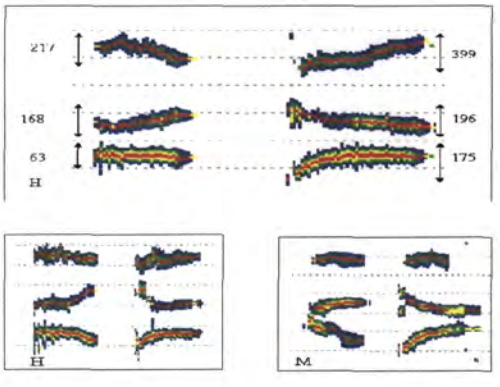
iki



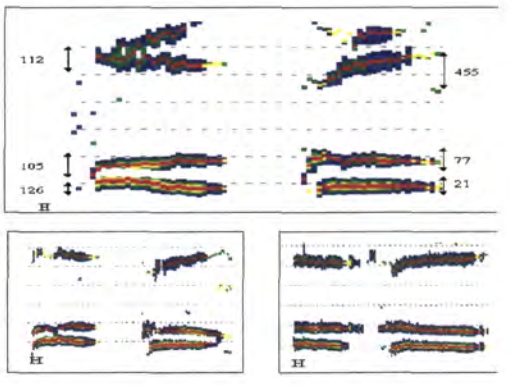
eke



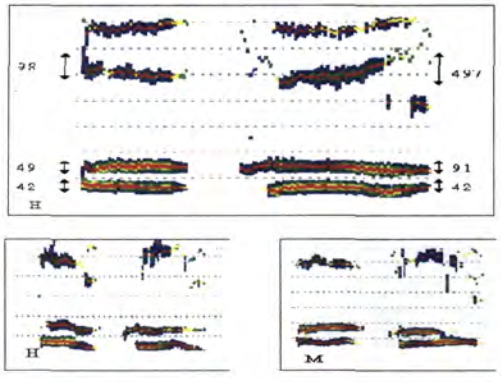
aka



oko



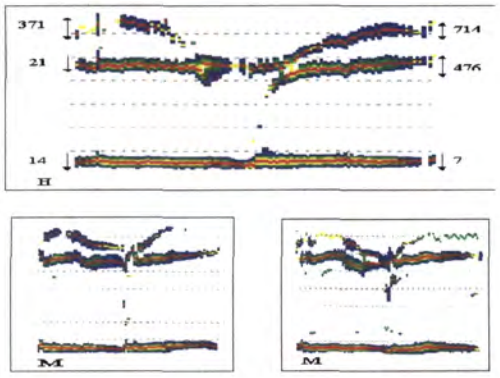
uku



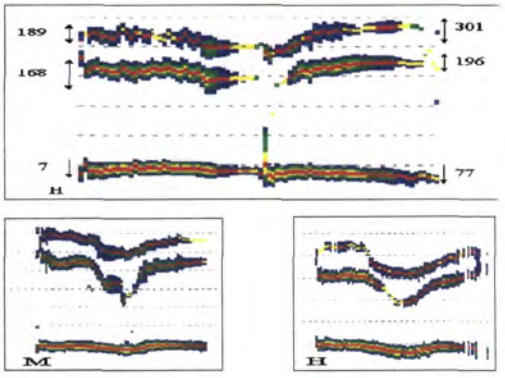
B APÉNDICE

[β]

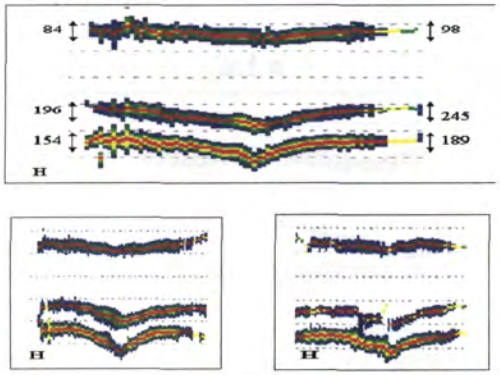
Iβi



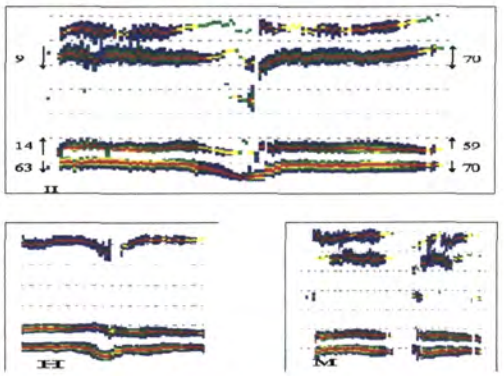
eβe



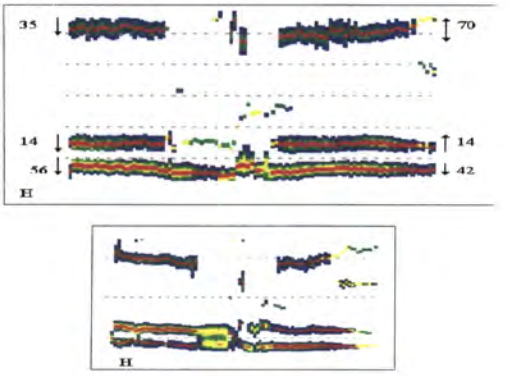
aβa



oβo

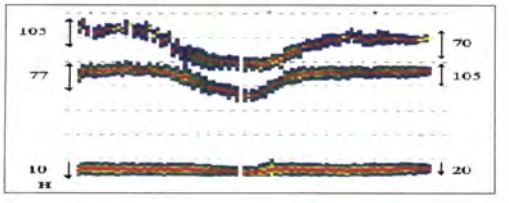


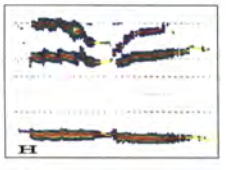
uβu

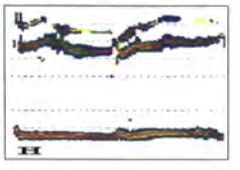


[d.]

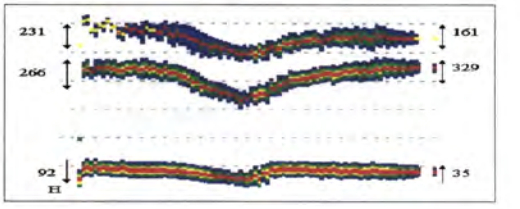
id.i

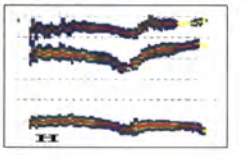


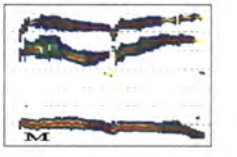




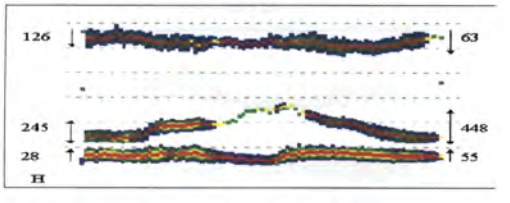
ed.e

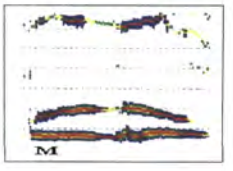


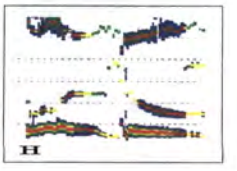




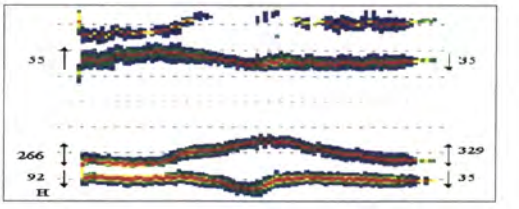
ad.a

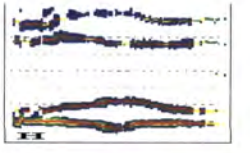


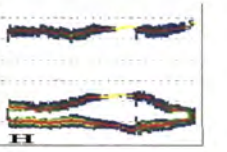




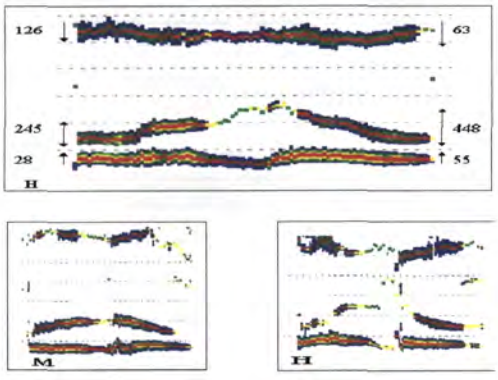
od.o





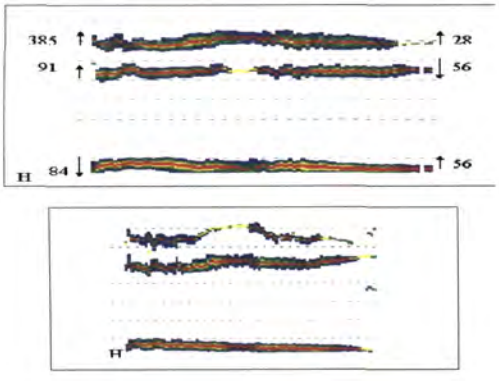


ud.u

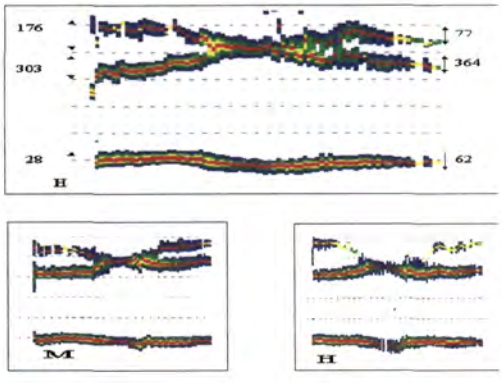


[γ]

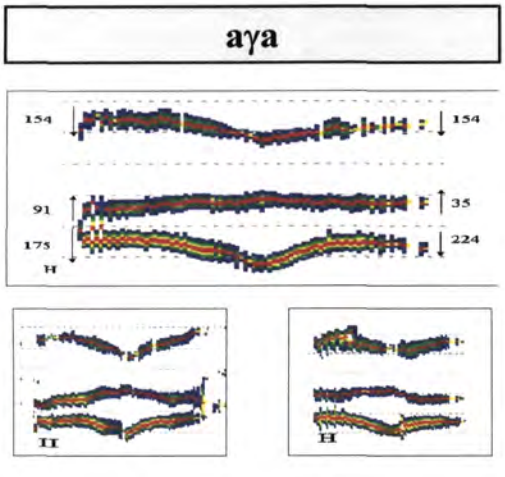
iγi



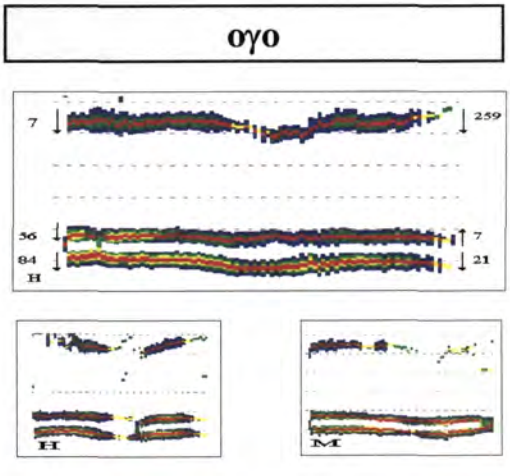
eγe



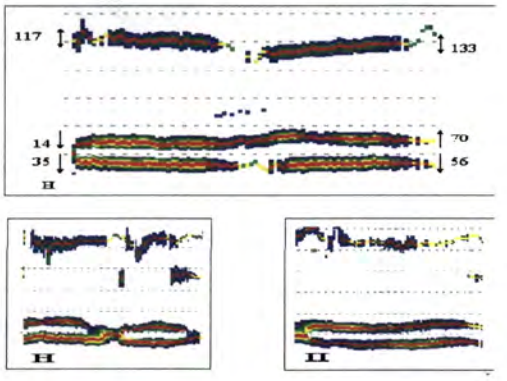
aγa



oγo



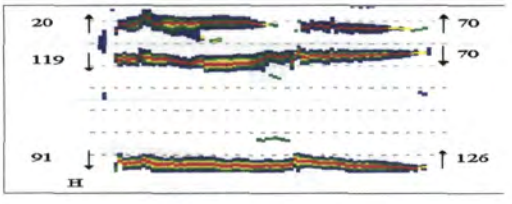
uγu

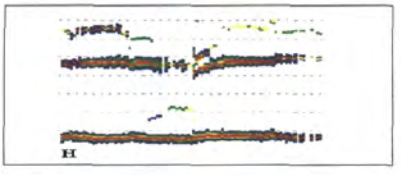


C APÉNDICE

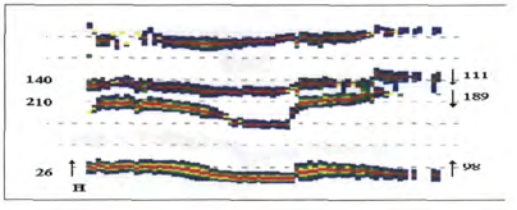
[m]

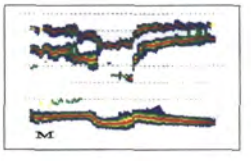
imi

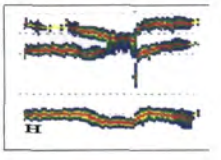




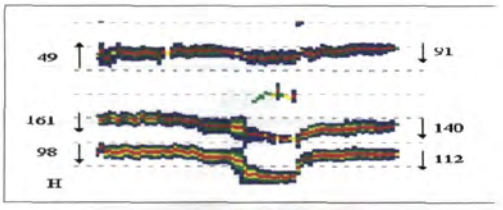
eme

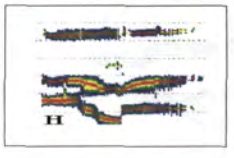


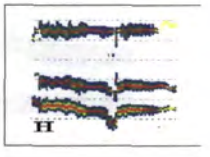




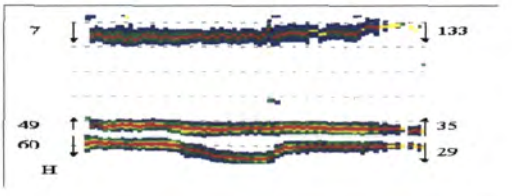
ama

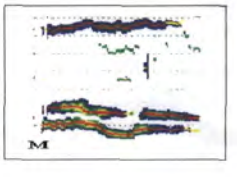


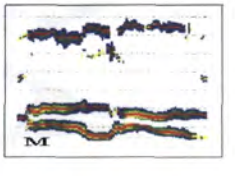




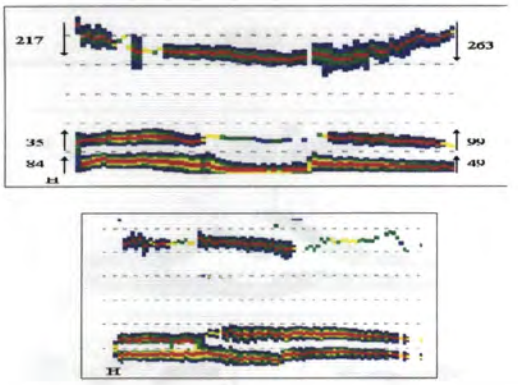
omo





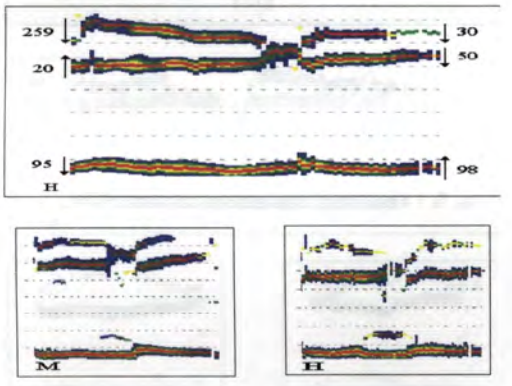


umu

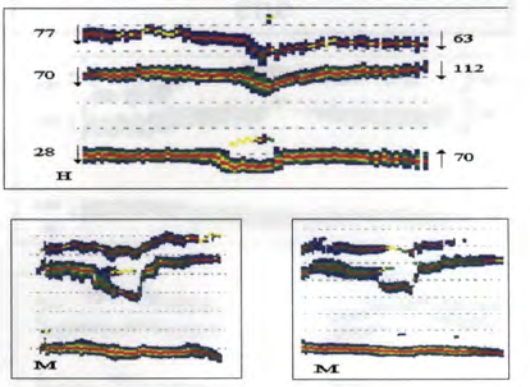


[n]

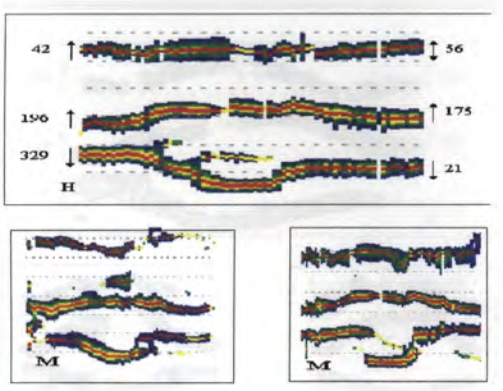
ini



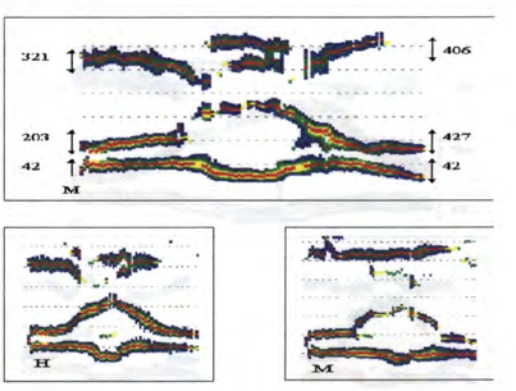
ene



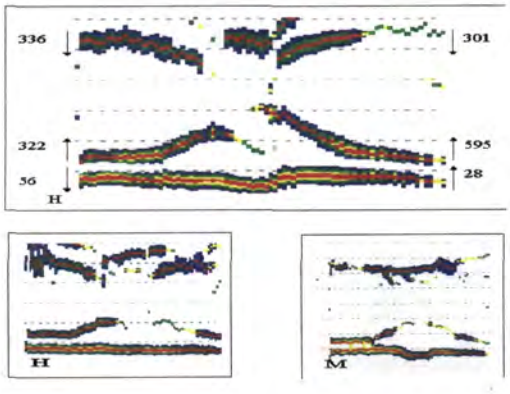
ana



ono

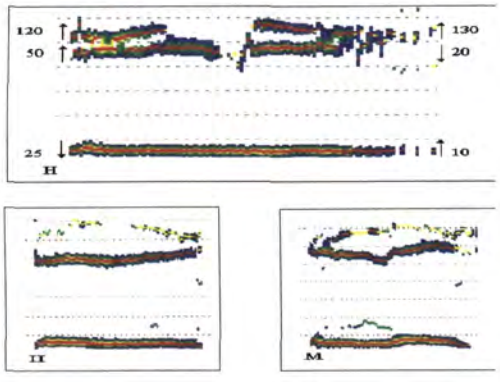


unu

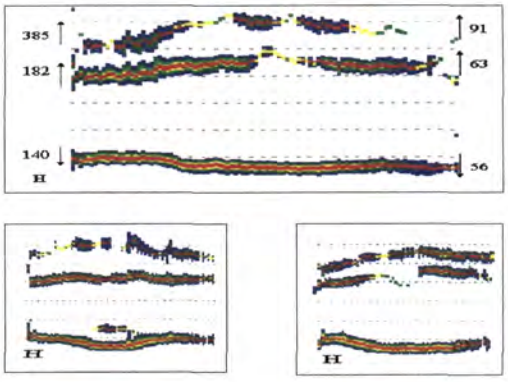


[ɳ]

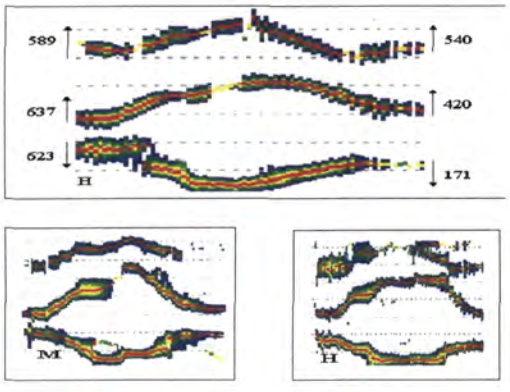
iɳi



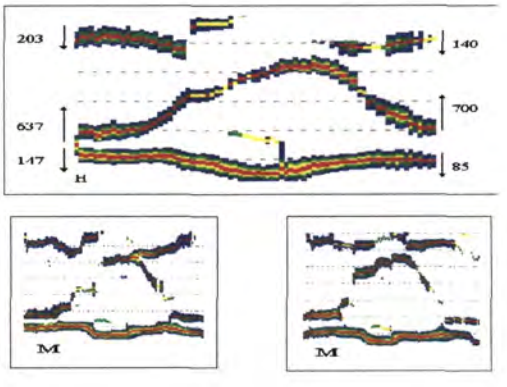
eɳe

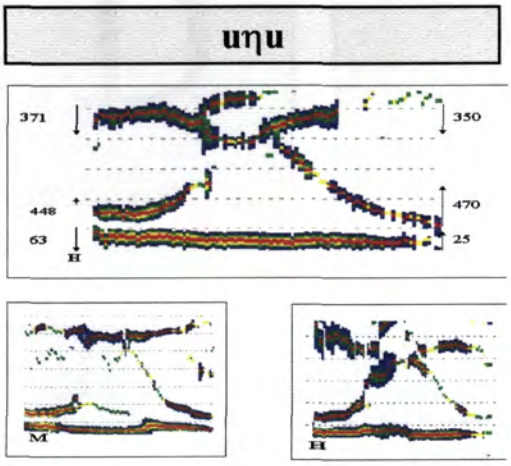


aɳa



oɳo

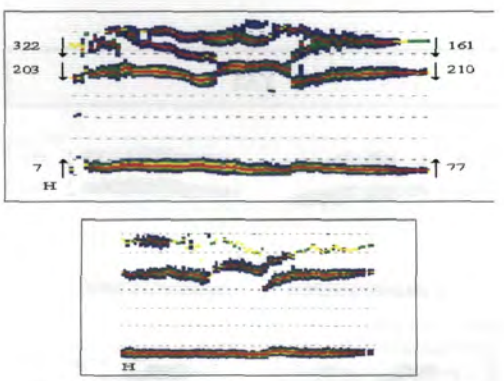



uηu
371
350
448
470
63
25
H
M
H

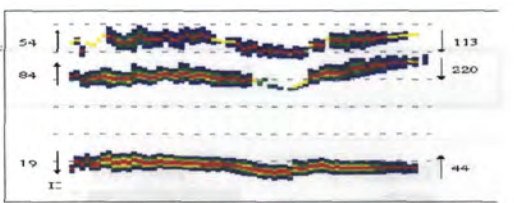
D APÉNDICE

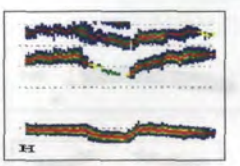
[I]

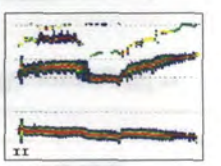
ili



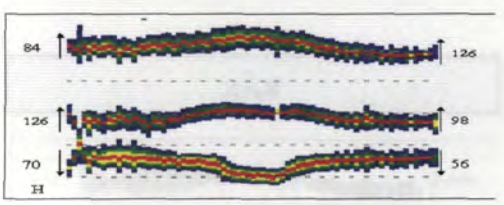
ele

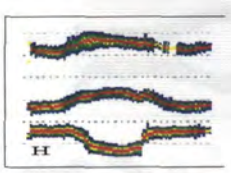


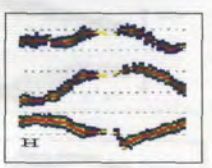




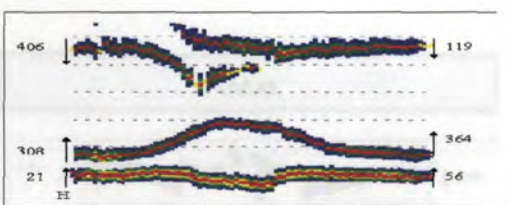
ala

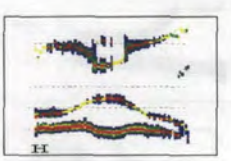


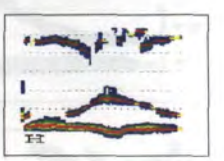




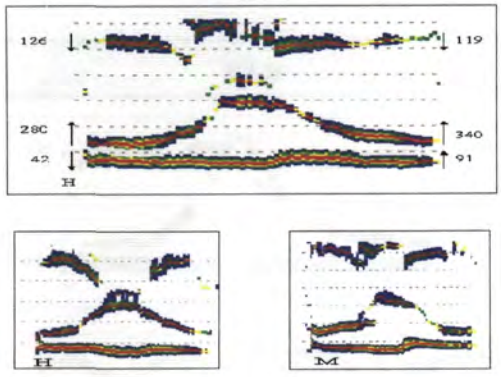
olo





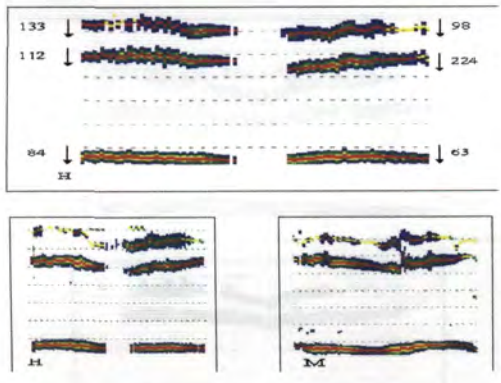


ulu

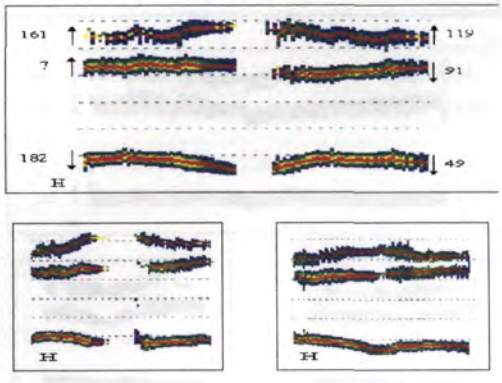


[λ]

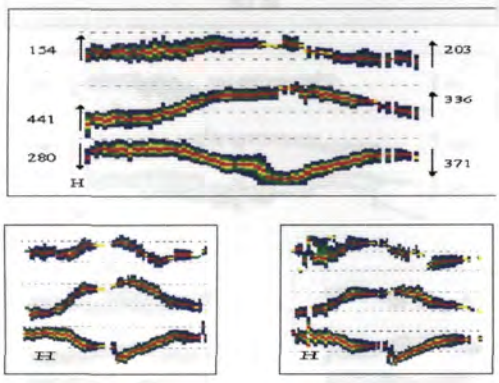
iλi



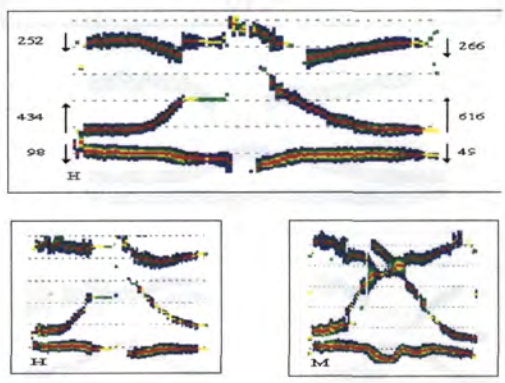
eλe



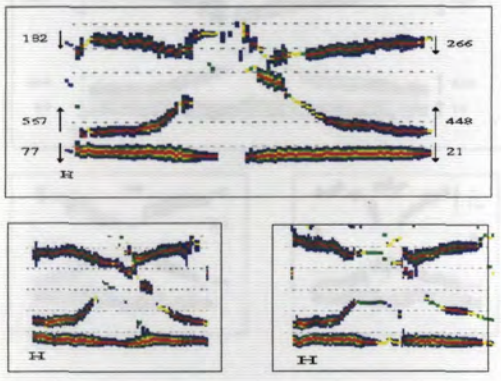
aλa



oλo

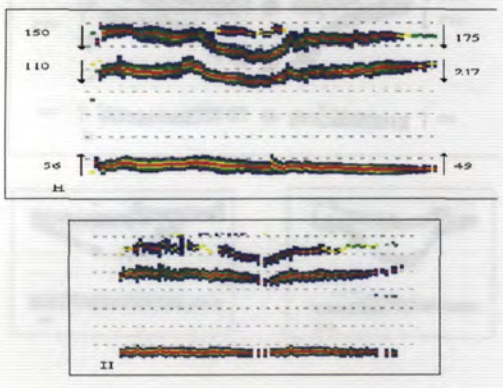


uλu

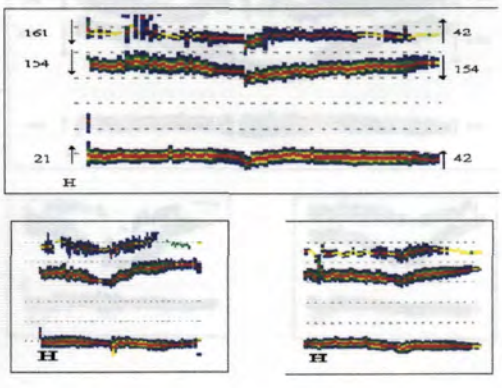


[r]

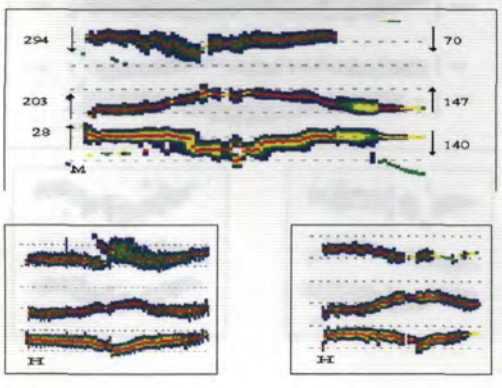
iri



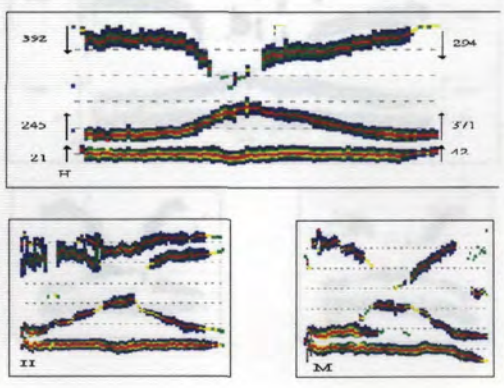
ere



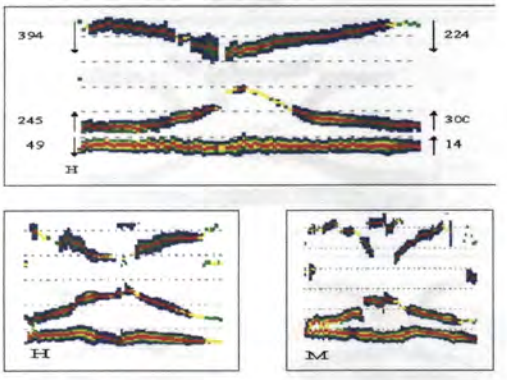
ara



oro

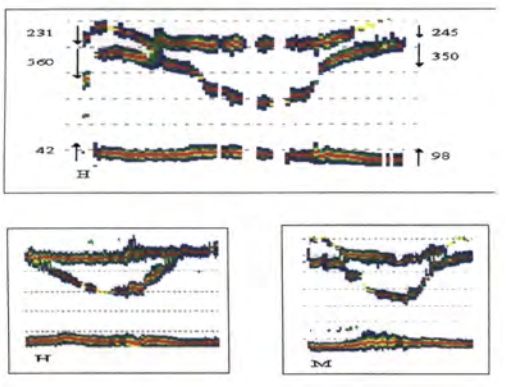


uru

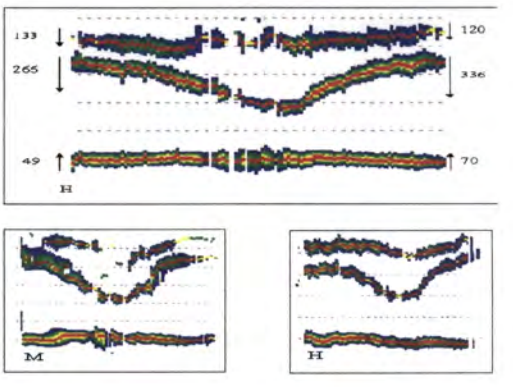


[rr]

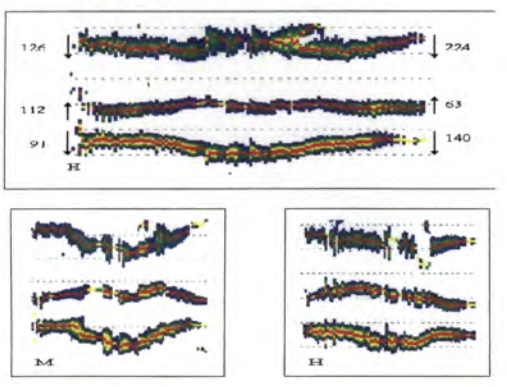
irri



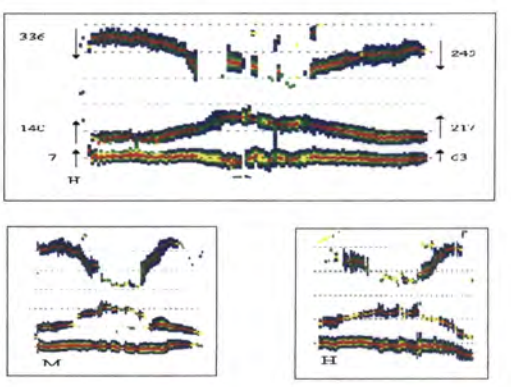
erre



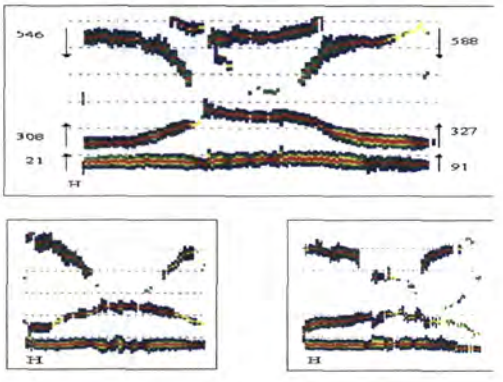
arra



orro



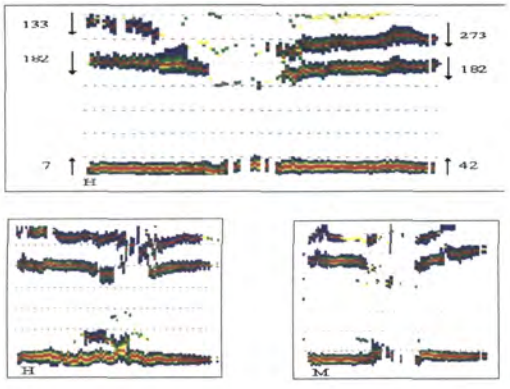
urru



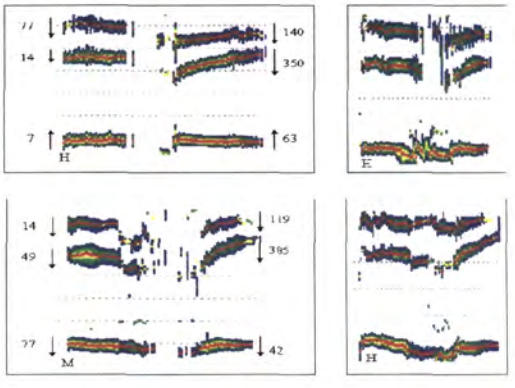
E APÉNDICE

[θ]

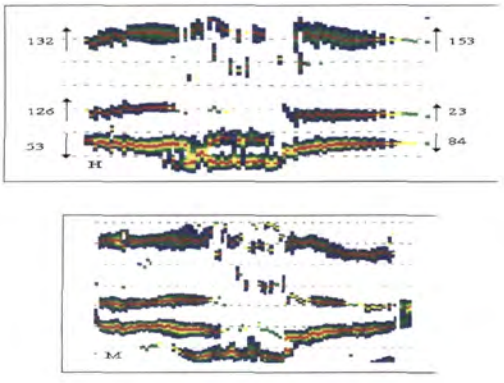
iθi



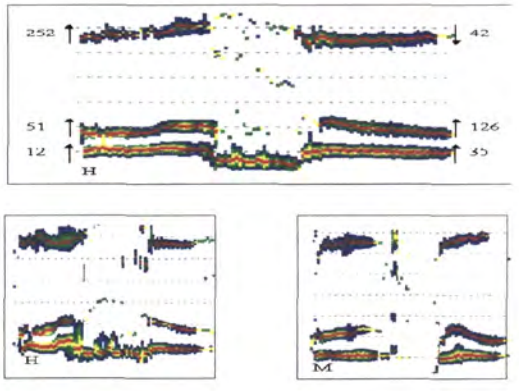
eθe



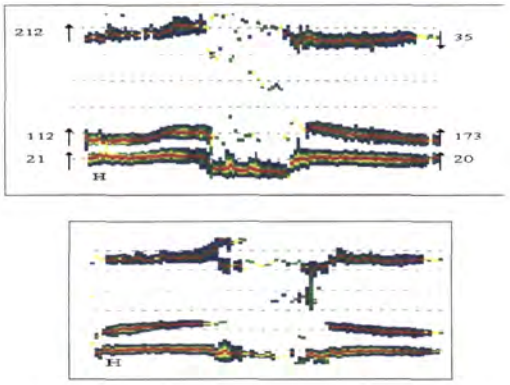
aθa



oθo

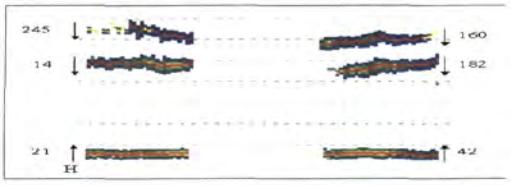


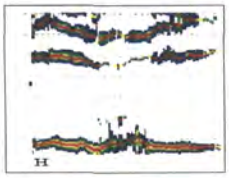
uθu

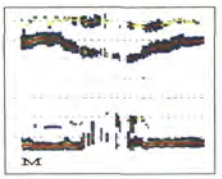


[s]

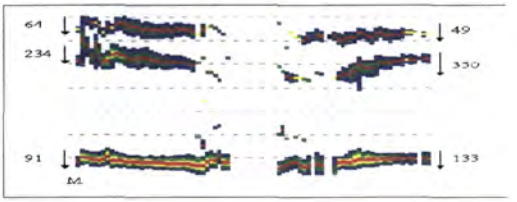
isi

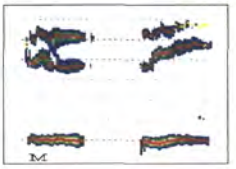


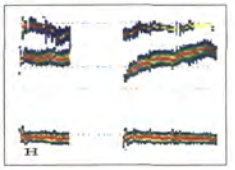




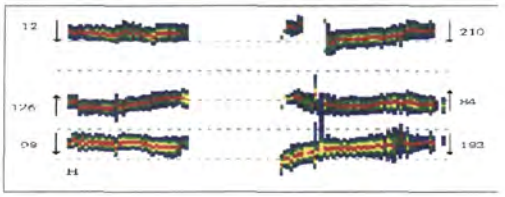
ese

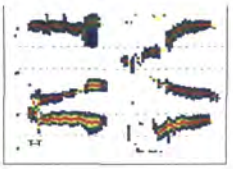


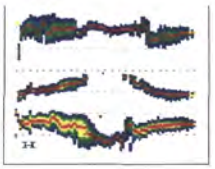




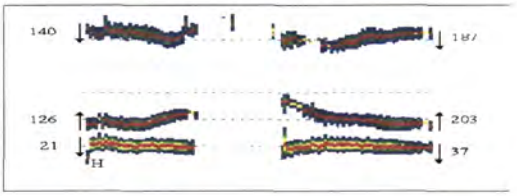
asa

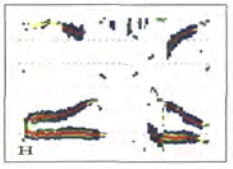


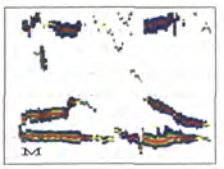




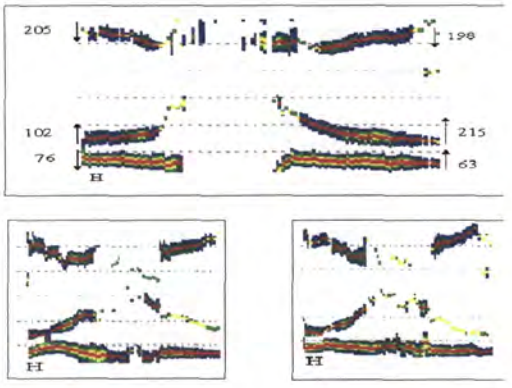
oso





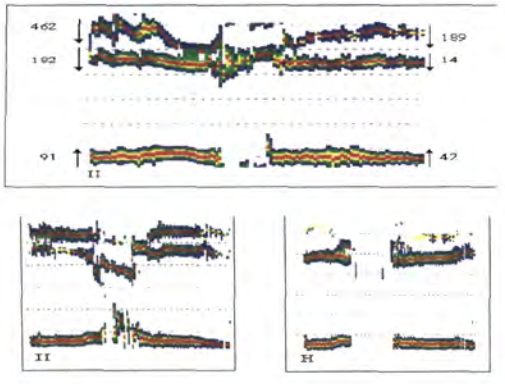


usu

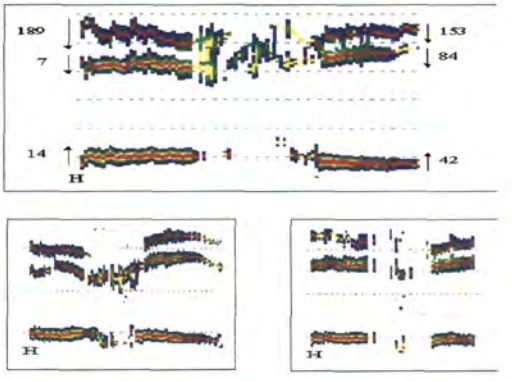


[x]

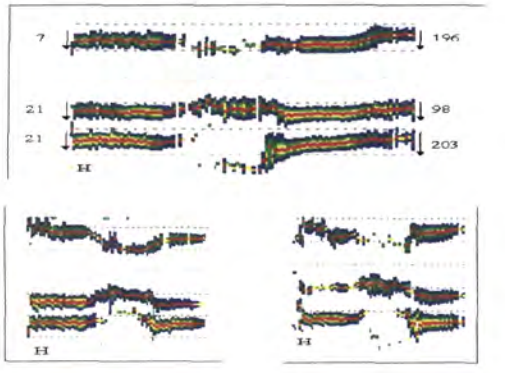
ixi



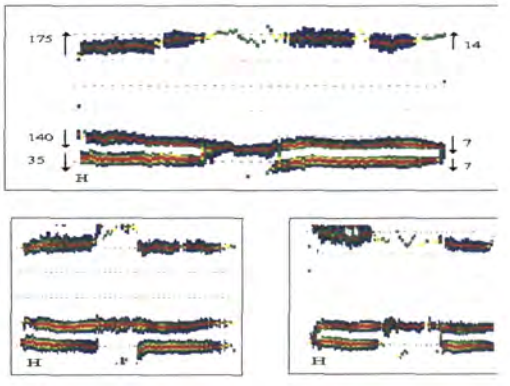
exe



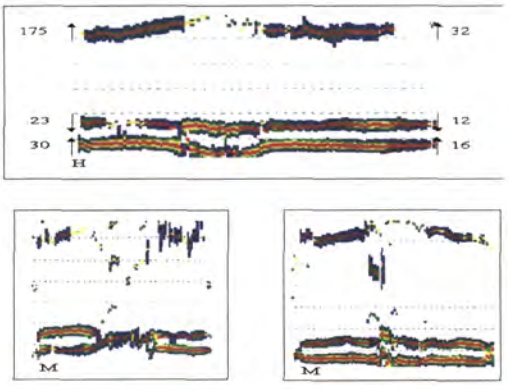
axa



oxo

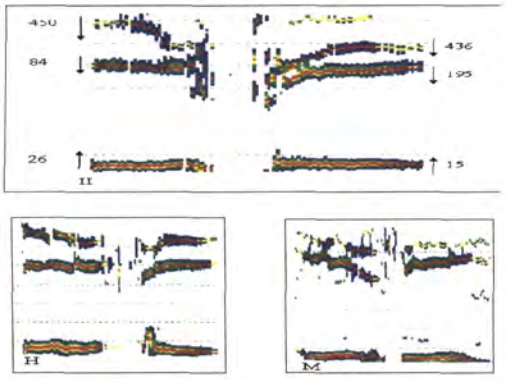


uxu

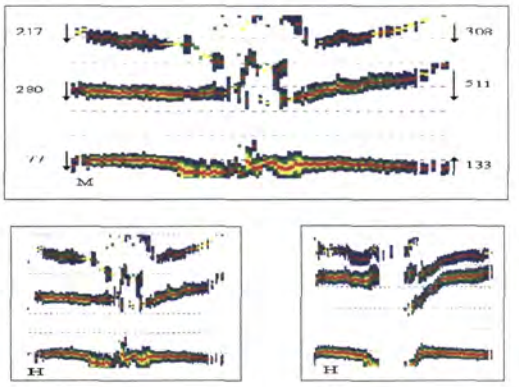


[f]

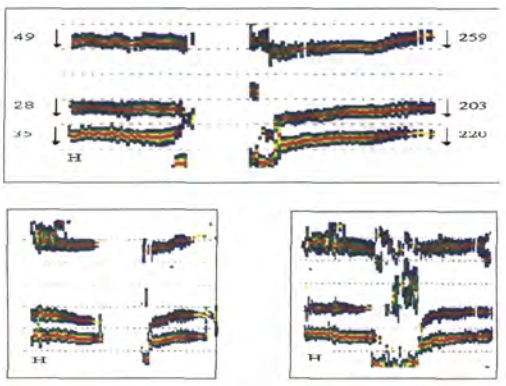
ifi



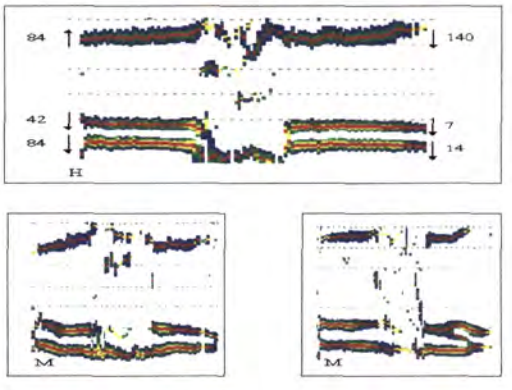
efe



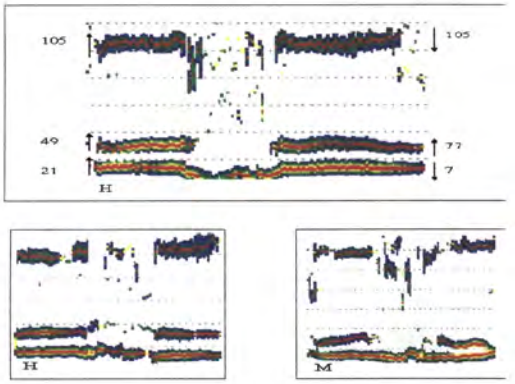
afa



ofo

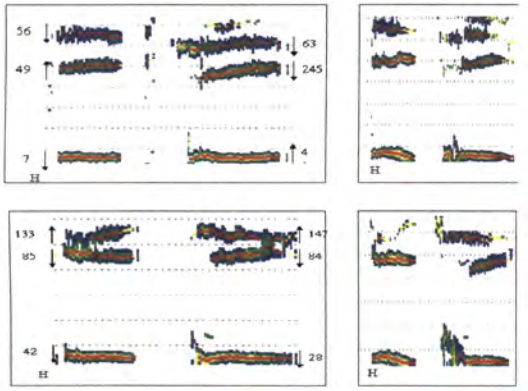


ufu

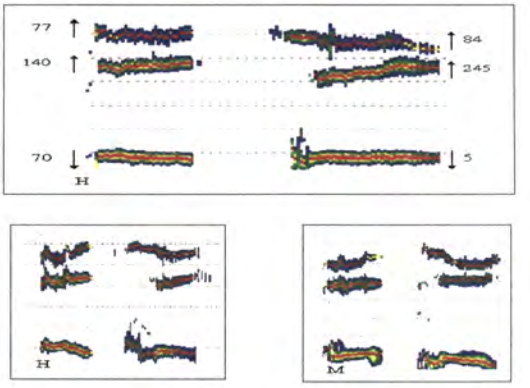


[tʃ]

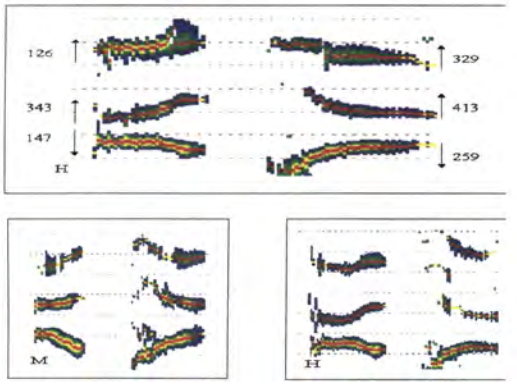
itʃi



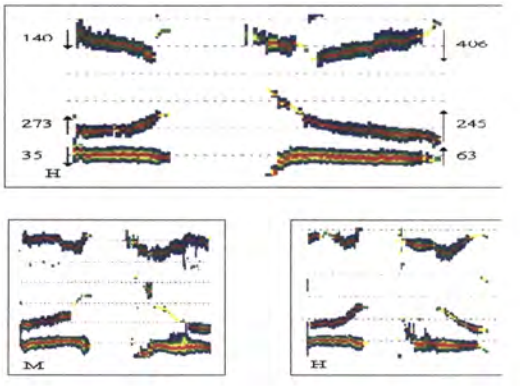
etʃe



atʃa



otʃo



utʃu