Guimarães - Portugal
acústica 2004

paper ID: 111 /p.1

Modelos Acústicos de Sílabas Consonante-Vocal para el Reconocimiento de Fricativas

S. Fernández[a] and S. Feijóo[b]

[a] *Department of Phonetics & Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom, santi@phon.ucl.ac.uk*
[b] *Departamento de Física Aplicada, Universidad de Santiago de Compostela, Facultad de Física, Campus Sur, 15782 Santiago de Compostela, España*

RESUMEN: La identificación de fonemas mejora si los oyentes disponen de parte de la señal adyacente. A nivel acústico-fonético, dos fuentes de información han sido propuestas como explicación para este efecto: a) la información coarticulatoria presente en segmentos adyacentes, y b) la identidad fonética de dichos segmentos adyacentes. En relación con este problema surge la cuestión de si el segmento correspondiente a un fonema y el contexto en el que se encuentra son procesados de forma conjunta o por el contrario se procesan de forma independiente y la información obtenida de cada segmento se combina en una segunda etapa. En este trabajo, se investiga el papel que tanto la información coarticulatoria presente en la vocal como la identidad de dicha vocal juegan en la identificación de consonantes fricativas. Se comparan, además, modelos acústicos que procesan los segmentos C y V conjunta o independientemente. Debido a las diferencias entre fricativas y vocales, diferentes parámetros acústicos han sido empleados. El mejor porcentaje de classificación de las fricativas se obtuvo cuando ambas fuentes de información vocálica fueron incluídas y los segmentos C y V se procesaron de forma conjunta (83% frente al 86% obtenido por los oyentes).

ABSTRACT: The perceptual identification of phonemes improves when listeners are presented with part of the signal surrounding a target phoneme. At the acoustic-phonetic level, two sources of information have been proposed in order to explain this effect: a) coarticulatory information in adjacent segments, and b) the phonetic identity of those adjacent segments. A related problem is whether the target and adjacent segments are processed jointly or whether they are processed independently and the evidence gathered from each segment is combined afterwards. In this paper, coarticulatory information in the vowel and the role of the phonetic identity of the vowel as cues for fricative recognition are investigated. Acoustic models that process the C and V segments either independently or jointly are compared. Several acoustic characterizations were tested due to the acoustic differences between fricatives and vowels. The best fricative recognition rates were achieved when both vocalic sources of information were included and the C and V segments were processed jointly (83% versus 86% listeners' identification rate).

1. INTRODUCCIÓN

Dos fuentes de información han sido propuestas en la literatura para explicar porqué la identificación de fonemas es mejor cuando los oyentes disponen de parte de la señal adyacente al segmento asociado con el fonema a identificar: a) información coarticulatoria en los segmentos adyacentes, y b) la identidad de dichos segmentos adyacentes.

La información coarticulatoria es debida a la dinámica del sistema de producción del habla. La coarticulación ayuda a preservar la continuidad en las propiedades acústicas de la señal hablada y, en consecuencia, proporciona información sobre la identidad de los fonemas adyacentes a un segmento dado de la señal. Por ejemplo, la transición entre consonantes y vocales contiene información sobre el lugar de articulación de la consonante.
Los efectos de la calidad de la vocal en la percepción del lugar de articulación de la consonante en sílabas CV son también notables de modo que Nearey otorga un importante papel a la identidad fonológica de dichos segmentos adyacentes a la hora de explicar los efectos que el contexto ejerce sobre la identificación de los fonemas [1]. En su modelo de percepción del habla, los oyentes evaluarían independientemente la información acústica en ambos segmentos. En una segunda etapa se tendría en cuenta la identidad fonológica de ambos antes de tomar una decisión final sobre la sílaba. En este sentido, la identificación de los segmentos adyacentes mediaría a la hora de interpretar la información acústica. van Son y Pols [2] encontraron que en sílabas CV la identificación de cada uno de los fonemas se ve afectada por la identificación que los oyentes hacen del fonema contigüo.
Dado que la información acústica empleada para la identificación de los fonemas se encuentra dispersa en la señal, surge la cuestión sobre la unidad fundamental del reconocimiento del habla. Nearey [1] propone unidades de tamaño acorde con los fonemas de las cuales se extrae información acústica que, en una segunda etapa, se combina para tomar una decisión final sobre la identidad de los segmentos. Por el contrario, van Son y Pols [2] observaron que las sílabas CV podrían ser identificadas de forma “integrada” y la identidad de los fonemas se derivaría a partir de dicha percepción. Los resultados para sílabas VC sugieren, sin embargo, unidades acordes con los fonemas. No está claro todavía si segmentos correspondientes a fonemas y su contexto han de ser procesados de forma separada o conjunta.
En un trabajo previo [3], la identificación de fricativas en estímulos C y CV mostró que los oyentes se benefician de la inclusión del segmento vocálico. Se construyeron modelos ocultos de Markov (MOM) para los dos tipos de estímulo. En la condición CV los modelos incluían información coarticulatoria presente en el segmento vocálico y procesaban el estímulo CV de forma “integrada”, como una sola unidad. Los resultados mostraron que los modelos acústicos se benefician de la inclusión de información coarticulatoria presente en el segmento vocálico. Conceptualmente el modelo es similar al de Nossair y Zahorian para consonantes plosivas [4]. Su modelo, sin embargo, emplea la trayectoria de los parámetros acústicos a lo largo del segmento CV, en lugar de un número fijo de estados (como hacen los MOM), para describir la dinámica del segmento CV. Usaron además un clasificador (análisis discriminante cuadrático (ADC)) más potente que el usado en los MOM (mezcla de gaussianas con matrices de covarianza diagonales). Esperamos que el modelo de Nossair y Zahorian obtenga mejores resultados que los MOM para este problema concreto.
En este trabajo, dichos análisis previos son aumentados mediante la inclusión de la identidad del segmento vocálico como fuente de información para la identificación de las fricativas. Además, modelos acústicos que procesan los segmentos C y V de forma independiente y, en una segunda etapa, combinan ambas fuentes de información para la identificación de las fricativas, son comparados con modelos que procesan el segmento CV de forma “integrada”. Los modelos se basan tanto en MOM, que constituyen el marco estadístico de los sistemas de reconocimiento automático del habla, como en la técnica desarrollada por Nossair y Zahorian.

2. MATERIALES Y MÉTODO

Los estímulos empleados en los experimentos son los presentados en [3]. Brevemente, sílabas fricativa-vocal formadas por /, , , , , / y /, , , , , / fueron pronunciadas por hablantes nativos de español (Galicia). Las señales fueron muestreadas a 32kHz y filtradas pasa-alta con una frecuencia de corte de 100Hz. Finalmente, se normalizaron para evitar diferencias notables en amplitud entre hablantes. Se construyeron dos tipos de estímulos: ruido fricativo aislado (condición C) y ruido fricativo aislado más 100ms de la vocal (condición CV). Las señales se separaron en dos conjuntos. El conjunto de entrenamiento se formó con 2800 estímulos = 5 fricativas × 5 vocales × (29 hablantes masculinos + 27 hablantes femeninos) × 2 repeticiones; y el conjunto de prueba con 500 estímulos = 5 fricativas × 5 vocales × (10 hablantes masculinos + 10 hablantes femeninos).

Veintiún oyentes participaron en los experimentos de percepción sobre el conjunto de prueba. El porcentaje de identificación de las fricativas fue 63.3% en la condición C y 86.0% en la condición CV. Para cada condición, se construyó un vector con el número de respuestas asignadas a cada estímulo. La correlación entre este vector y otro similar formado por las probabilidades *a posteriori* asignadas a cada estímulo por los modelos acústicos se usará como indicador de la validez de dichos modelos. Los oyentes fueron divididos en dos grupos y la correlación entre sus respuestas fue calculada: r=0.91 en la condición C y r=0.97 en la condición CV. Estos coeficientes indican cuál es el valor máximo esperado para la correlación entre los modelos acústicos y las respuestas de los oyentes. Un análisis más exhaustivo de los experimentos puede encontrarse en [3].

Resulta difícil saber *a priori* qué parámetros acústicos y qué orden es mejor para representar las características de un segmento CV debido a las diferencias acústicas entre consonantes y vocales. Los siguientes parámetros acústicos se usaron para los modelos basados en el ADC: coeficientes cepstrum obtenidos mediante LPC, cepstrum en escala lineal, cepstrum en escala Mel y banco de filtros en escala Mel. Las señales fueron remuestreadas a 20kHz para disponer de un número de coeficientes manejable (20 para el cepstrum y 27 para el banco de filtros). Los ADC se llevaron a cabo para todos los ordenes de la parametrización.

Los parámetros acústicos se calcularon sobre ventanas separadas. También se analizaron los modelos empleando la trayectoria de dichos parámetros a lo largo del segmento. En base a análisis previos, en el primer caso tres ventanas de 40ms se usaron para el ruido fricativo: una al comienzo, otra en el medio y otra al final del ruido; y dos ventanas de 20ms se usaron para la vocal: una al comienzo y otra al final del segmento vocálico. En el segundo caso, los parámetros se calcularon sobre ventanas de 15ms, solapadas 10ms, que cubrían el segmento C, V o CV. Finalmente la trayectoria de dichos parámetros a lo largo del segmento fue codificada mediante los tres primeros coeficientes de una serie coseno [4].

El paquete HTK (http://htk.eng.cam.ac.uk) se usó para los análisis con MOM. Se estudiaron modelos con cuatro y ocho gaussianas. Las probabilidades de observación se modelaron con matrices de covarianza diagonales. Se usaron modelos de izquierda-a-derecha con tres estados para el segmento C, dos estados para el segmento V y cinco estados para el segmento CV.

3. MODELOS ACÚSTICOS

El objetivo de los modelos es identificar la fricativa en las condiciones C y CV. Si el segmento C se clasifica como una de las cinco fricativas (modelo C), no se incluye información sobre la identidad del segmento vocálico. Por el contrario, si se emplean veinti-cinco grupos de clasificación (5 fricativas × 5 vocales) incluimos información sobre la identidad de la vocal (modelo Cq). Análogamente, el segmento CV puede clasificarse como una de las cinco fricativas (modelo CV) o pueden usarse veinticinco grupos de clasificación (modelo CVq). En el primer caso se incluye información coarticulatoria presente en la vocal mientras que en el segundo caso se incluye además información sobre la identidad del segmento vocálico.
Si los segmentos C y V se procesan de forma independiente y la información obtenida se combina en una etapa posterior, surgen varias posibilidades. Los segmentos C y V son, primeramente, clasificados en cinco o veinticinco grupos dependiendo de si sólo se emplea información coarticulatoria o también la identidad de la vocal se tiene en cuenta (modelos C-V y C-Vq, respectivamente). En segundo lugar, las probabilidades obtenidas para cada segmento son combinadas. Tres funciones se han estudiado: función O, Y e Y ponderada. La función O implica que la decisión final se basa en sólo uno de los segmentos, aquel que contiene la mayor información sobre la identidad de la fricativa:

$$P_{CV} = P_c + P_V - P_C \cdot P_V \qquad (1)$$

La función Y implica que ambos segmentos son tenidos en cuenta:

$$P_{CV} = P_C \cdot P_V \qquad (2)$$

La función Y ponderada se emplea para relajar la condición de independencia asumida en la ecuación (2):

$$P_{CV} = \left(P_c\right)^{w_C} \cdot \left(P_V\right)^{w_V} \qquad (3)$$

Los pesos w_C y w_V se estiman a partir del conjunto de entrenamiento. Los estímulos del conjunto de entrenamiento se clasificaron con el método de *dejar-uno-fuera* y se obtienen así valores estimados de P_C y P_V. A partir de unos pesos iniciales iguales a la unidad y mediante el algorítmo de descenso del gradiente, se estiman los valores de los pesos que verifican el criterio de mínimo error de clasificación.
Notar que, cuando se emplean veinticinco grupos de clasificación, la etapa final implica sumar las probabilidades asignadas a, por ejemplo, /● ●, ● ●, ● ●, ● ● , ● ● / para obtener la probabilidad de que un determinado estímulo sea clasificado como la fricativa /● /. Esto permite calcular la correlación con las respuestas de los oyentes.

3.1 Resultados

Los resultados de clasificación empleando el ADC se muestran en la figura 1 para cada uno de los parámetros acústicos usando bien la trayectoria de los mismos a lo largo del segmento o bien un número fijo de ventanas separadas. Los modelos para el segmento CV incluyen información coarticulatoria presente en la vocal, pero no información sobre la identidad de la vocal. No obstante, los resultados son mejores cuando se usa el segmento CV que cuando se usa sólo el segmento C. En general, el orden de los modelos según incrementa el porcentaje de clasificación es: C, O, Y, Y ponderada, CV. El orden de la caracterización acústica no ha de ser necesariamente alto: 8-12 coeficientes son suficientes. De los resultados para el banco de filtros puede observarse que incluir información por encima de 4.5kHz no incrementa de forma significativa los porcentajes de clasificación.

La figura 2 muestra la correlación entre la clasificación acústica y las respuestas de los oyentes. El patrón es similar al de los porcentajes de clasificación, aunque la mejor correlación no coincide necesariamente con el mejor porcentaje de clasificación. La correlación es superior para el modelo CV cuando se emplea la trayectoria de los parámetros. Cuando se usan ventanas separadas, los coeficientes de correlación son superiores para las funciones Y e Y ponderada. Este resultado muestra la importancia de modelar la dinámica de los parámetros acústicos de un modo más preciso que el empleado con ventanas separadas. Los porcentajes de clasificación no muestran este efecto claramente.

En general, las mayores diferencias dependen del uso de información presente en el segmento vocálico. Los resultados con MOM usando sólo información coarticulatoria o, además, la identidad del segmento vocálico se muestran en la tabla 1. Se muestran también los resultados con el ADC empleando la trayectoria de los coeficientes cepstrum en escala Mel para comparar ambos métodos. El ADC no se llevó a cabo incluyendo información sobre la identidad de la vocal porque el conjunto de entrenamiento no es suficicientemente grande como para estimar con fiabilidad matrices de covarianza completas.

Los resultados muestran que los modelos se benefician de la inclusión de ambas las fuentes de información presentes en la vocal. Los porcentajes de clasificación en la condición C son mayores que los de los oyentes. Lo contrario ocurre en la condición CV. A pesar de incluir

Tabla 1 – *Resultados de clasificación y correlación para los diferentes modelos usando cepstrum en escala Mel. En los modelos C-V y C-Vq se empleó la función Y. Los porcentajes obtenidos por los oyentes en las condiciones C y CV fueron 63.3% y 86.0%, respectivamente. La consistencia inter-oyente fue 0.91 en la condición C y 0.97 en la condición CV.*

Condición	MOM		ADC	
	%	*r*	%	*r*
C	80.4	0.77	72.0	0.67
Cq	81.8	0.76	NA	
C-V	79.4	0.80	78.8	0.78
C-Vq	80.4	0.80	NA	
CV	81.2	0.81	82.8	0.79
CVq	83.0	0.84	NA	

información sobre la dinámica de los parámetros en el ADC, los resultados son mejores con los MOM. El modelo acústico con los mejores resultados incluye información coarticulatoria, sobre la identidad de la vocal, y procesa ambos segmentos de forma conjunta. El porcentaje de clasificación en este caso (83%) es ligeramente inferior al de los oyentes (86%). La diferencia entre la correlación (r=0.84) y la consistencia inter-oyente (r=0.97) es más notable.

4. CONCLUSIÓN

Varios modelos acústicos para el reconocimiento de fricativas en sílabas CV han sido propuestos. Los porcentajes de clasificación mejoran cuando se incluye información presente en el segmento vocálico. Los modelos se benefician no sólo de la información coarticulatoria presente en la vocal sino que también se benefician de información sobre la identidad del segmento vocálico, tal y como proponen algunas de las teorías sobre la percepción fonética.
En nuestro caso, un procesado conjunto de los segmentos C y V da los mejores resultados (83%), comparables al porcentaje de identificación obtenido por los oyentes (86%). Sin embargo, es necesario puntualizar que la correlación entre la clasificación automática y la identificación perceptual (r=0.84) es baja comparada con la consistencia inter-oyente (r=0.97). Es todavía necesario mejorar el modelado acústico de los sistemas de reconocimiento automático.

RECONOCIMIENTOS

S. Fernández ha sido financiado mediante una beca Marie Curie del programa de la Comunidad Europea “Improving the Human Research Potential and the Socio-Economic Knowledge Base” bajo contrato número HPMF-CT-2002-02129.

REFERENCIAS

[1] T. M. Nearey; *Context effects in a double-weak theory of speech perception*. Language and Speech, vol. 35, págs. 153-171, 1992.

[2] R. J. J. H. van Son y L. C. W. Pols; *Perisegmental speech improves consonant and vowel identification.* Speech Communication, vol. 29, págs. 1-22, 1999.

[3] S. Fernández y S. Feijóo; *Comparing HMM-based recognition to perceptual phonetic integration in fricative-vowel syllables*. In Proceedings of the 15th International Congress of Phonetic Sciences, págs. 1433-1436, Barcelona, 2003.

[4] Z. B. Nossair y S. A. Zahorian; *Dynamic spectral shape features as acoustic correlates for initial stop consonants.* Journal of the Acoustical Society of America, vol. 89, págs. 2978-2991, 1991.

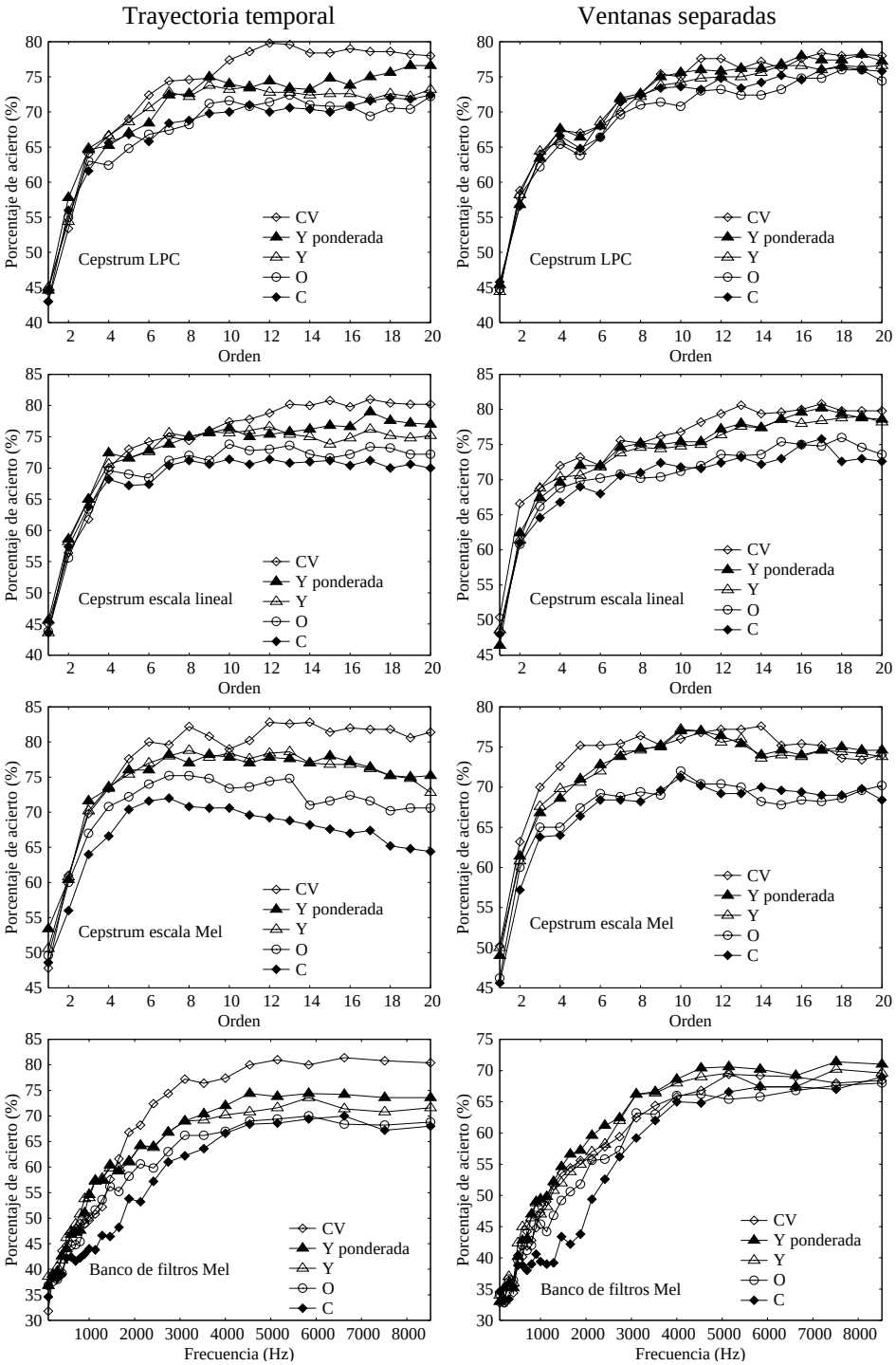

Figura 1 – *Resultados de clasificación en las condiciones C y CV usando el ADC. Para la condición CV, varios modelos han sido usados para integrar la información presente en los segmentos C y V (ver texto). Los resultados se muestran para diferentes parámetros acústicos. A la izquierda se muestran los análisis empleando la trayectoria de los parámetors a lo largo del segmento, a la derecha se muestran los resultados usando ventanas separadas.*

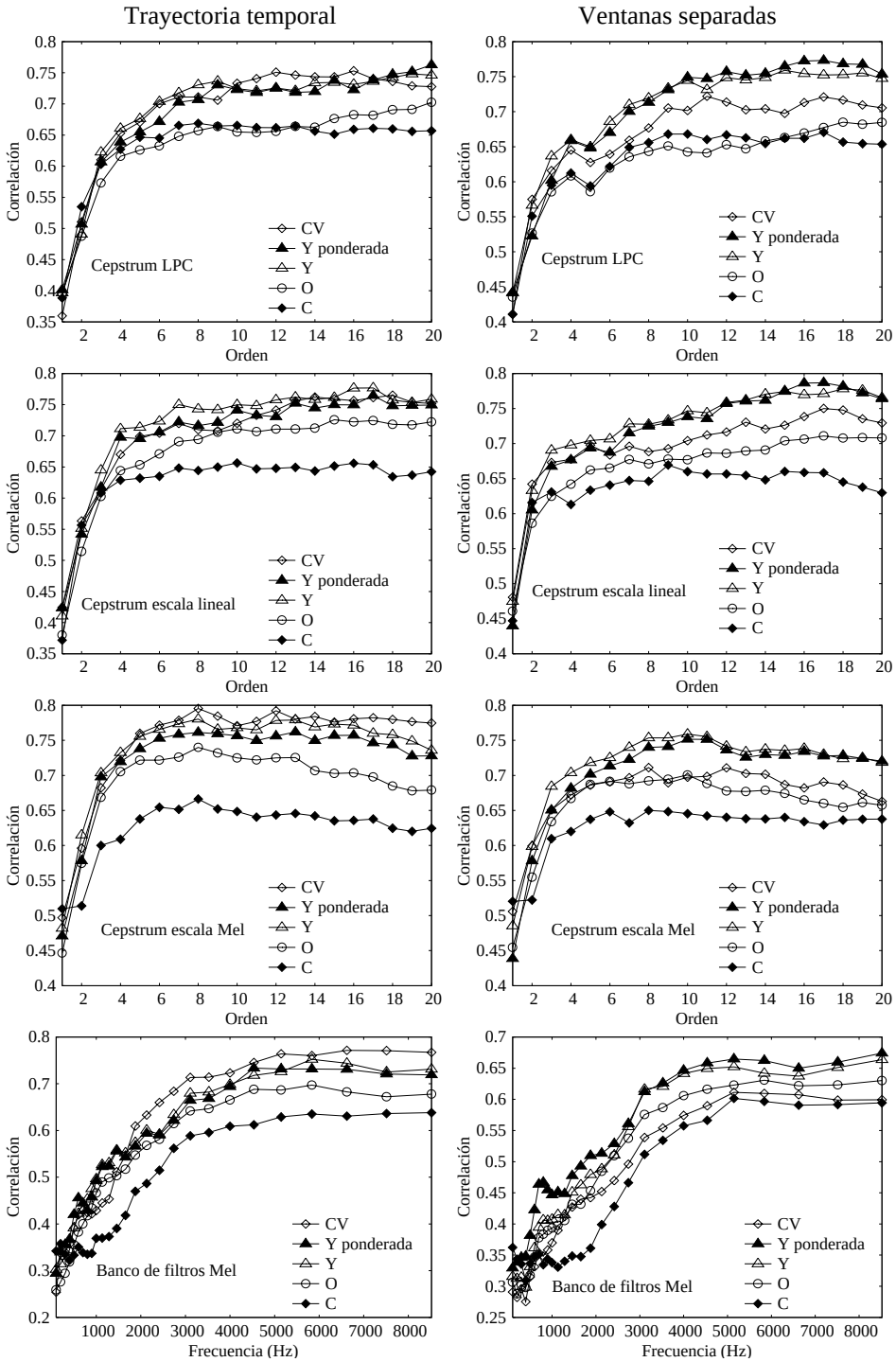

Figura 2 – *Correlación entre los oyentes y los modelos acústicos en las condiciones C y CV (ADC). En la condición CV, varios modelos han sido usados para integrar la información de los segmentos C y V (ver texto). Los resultados se muestran para diferentes parámetros acústicos. A la izquierda se muestran los análisis empleando la trayectoria de los parámetors a lo largo del segmento, a la derecha se muestran los resultados usando ventanas separadas.*