Evaluación del uso de cocleagramas para la identificación del hablante en fonética forense

Evaluation of cochleograms for the speaker identification in forensic Phonetics

Jordi Cicres

Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, ForensicLab – Laboratori de Lingüística Forense Jordi, Roc Boronat 138, 08018, Barcelona, España.
e-mail: cicres@upf.edu

El presente estudio evalúa la posibilidad de utilizar cocleagramas (representación gráfica de los sonidos que representa la excitación de la membrana basilar por unidad de tiempo, y refleja en cierto modo la audición humana) como herramienta para identificar hablantes con fines forenses. Los resultados (a partir de la comparación de cocleagramas mediante coeficientes de correlación) muestran una tendencia global discriminante (la variación intra-hablante es menor que la inter-hablante), aunque de modo desigual. Asimismo, se constata que se obtienen mejores resultados en la comparación de sonidos dubitativos que en palabras sueltas.

Palabras clave: evaluación de los cocleagramas para la identificación del hablante.

The present study evaluates the possibility to use cochleagrams (a graphic representation of the sounds that represents the excitation pattern of the basilar membrane by unity of time, and reflect in some way the human audition) as a tool to identify speakers. Results from the comparison of cochleagrams by means of Pearson's correlation coefficients show a global discriminating trend (intra-speaker variation is minor that inter-speaker variation). Likewise, it is shown that better results are obtained in the comparison of hesitation sounds than in single words.

Keywords: Evaluation of cochleograms for speaker identification.

1. INTRODUCCIÓN

Una de las tareas encomendadas a los fonetistas forenses es la identificación de hablantes con el fin de aportar una opinión técnica en un proceso judicial acerca de si el autor de una grabación dubitada (es decir, cuya autoría es desconocida por el juez o tribunal) se corresponde al autor de una o varias grabaciones indubitadas (cuya autoría no está cuestionada). Para ello, los fonetistas forenses utilizan dos conjuntos de técnicas complementarias: la lingüística y la fonético-acústica. La primera consiste en identificar rasgos fonéticos, fonológicos, morfológicos, sintácticos, pragmáticos, semánticos y discursivos (es decir, rasgos lingüísticos) a partir de la escucha sistemática (por el experto) de las grabaciones dubitadas e indubitadas. En este caso, los lingüistas

aplican sus conocimientos en dichas áreas lingüísticas con el fin de identificar aquellos aspectos coincidentes o divergentes en las grabaciones analizadas, los cuales aportarán indicios acerca de la autoría de la grabación dubitada. La segunda técnica (la fonético-acústica) se fundamenta en el análisis y comparación cualitativa y cuantitativa de un conjunto de variables acústicas mediante las representaciones gráficas de los sonidos (básicamente espectros, espectrogramas, oscilogramas, mapas de formantes, cepstrums, y líneas de frecuencia fundamental e intensidad).

Las variables tradicionalmente analizadas tienen que ver no sólo con las características físicas de los sonidos sino también con las elecciones lingüísticas que conforman el idiolecto del hablante, término que se puede definir como la elección individual de cada hablante de algunas formas lingüísticas (fonéticas, fonológicas, morfológicas, sintácticas, pragmáticas, discursivas) determinadas que le ofrece la lengua [1]. Por tanto, el idiolecto es –por lo menos teóricamente– individual y único.

En el plano acústico, en la actualidad existen dos grupos de expertos: por un lado, los que utilizan técnicas no automáticas, basadas en el análisis de rasgos fonéticos concretos, y por el otro, los laboratorios que utilizan técnicas automáticas o semiautomáticas basadas en ratios de similitud (*likelihood ratios, LR*) y modelos de mezclas gaussianas (*Gaussian Mixture Models, GMMs*).

El análisis lingüístico sólo es posible mediante el método auditivo (hasta la actualidad no existen métodos automáticos que analicen realizaciones de variables fonológicas, morfológicas, sintácticas, etc.). Sin embargo, el análisis fonético-acústico (es decir, la descripción de las peculiaridades de los sonidos del habla) puede realizarse utilizando herramientas de análisis acústico, mediante la audición o mediante una combinación de ambas. En la actualidad, la mayoría de laboratorios utilizan complementariamente el análisis auditivo y el análisis acústico. En general se obtienen resultados satisfactorios [2].

Por otro lado, los humanos tenemos cierta capacidad innata de identificar hablantes únicamente mediante la audición, aunque con limitaciones evidentes. En las décadas de 1930 y 1940, F. McGehee ([3] y [4]) realizó los primeros estudios científicos acerca de la habilidad de las personas no expertas para identificar hablantes solamente mediante la audición de las muestras, poniendo especial atención al tiempo transcurrido desde el primer contacto hasta la tarea de identificación. Más recientemente, otros autores han continuado las investigaciones controlando otros factores, tales como la familiaridad con las voces [5], [7]; la calidad de la grabación [5], [6]; el canal de transmisión [5] y [7]; la familiaridad con la voz dubitada [5], [7], el estilo de habla [7]; el lapso de tiempo entre las grabaciones comparadas [3], [4], [5], [6]; y el nivel de entrenamiento o conocimientos fonéticos de los oyentes [9]. En todos los estudios se detectan serias limitaciones de la habilidad innata, por lo que su uso en el ámbito judicial debe de ser considerado con mucha cautela.

Sin embargo, en las identificaciones expertas puede ser útil aprovechar esta capacidad humana innata. Para ello puede resultar útil utilizar los sistemas de representación del sonido que emulan la audición humana (los cocleagramas). En este artículo realizamos comparaciones de la señal basadas en el coeficiente de correlación entre cocleagramas y cuantificamos dichas diferencias.

El objetivo de este artículo es doble: en primer lugar, evaluar el poder discriminante de los cocleagramas –como herramienta complementaria al análisis lingüístico y acústico– para la identificación forense de hablantes; y en segundo lugar, establecer límites en el uso de técnicas que emulan la audición humana en fonética forense.

2. LOS COCLEAGRAMAS

Los sonidos del habla son complejos: presentan información acústica en un amplio rango de frecuencias a lo largo del tiempo. Para reflejar esta riqueza en el análisis fonético-acústico del habla son necesarios gráficos que permitan representar la estructura espectral (la energía presente en cada rango de frecuencias) de los sonidos.

En fonética forense, la identificación de hablantes se basa en una parte importante en la comparación de gráficos que representan dicha distribución. Para representar la estructura espectral se pueden utilizar varios tipos de gráficos (espectros FFT, LPC o LTAS, espectrogramas, cocleagramas, etc.). Únicamente los espectrogramas y los cocleagramas muestran el eje temporal..

La principal diferencia entre los espectrogramas y los cocleagramas es que los primeros miden la densidad de energía en cada frecuencia mientras que los segundos reflejan el grado de excitación de la membrana basilar. En ambos casos se representa el tiempo. La tabla 1 muestra las principales diferencias entre ambos tipos de gráficos, y las figuras 1 y 2 muestran un ejemplo de la misma secuencia fonética [b↔Pitát] representada mediante un espectrograma y un cocleagrama respectivamente:

Tabla 1

Resumen de las diferencias entre espectrogramas y cocleagramas.

	Espectrograma	**Cocleagrama**
Unidad temporal	Segundos	Segundos
Unidad de frecuencia	Hz	Bark
Análisis	Densidad de energía acústica en cada frecuencia por unidad de tiempo	Excitación de la membrana basilar por unidad de tiempo

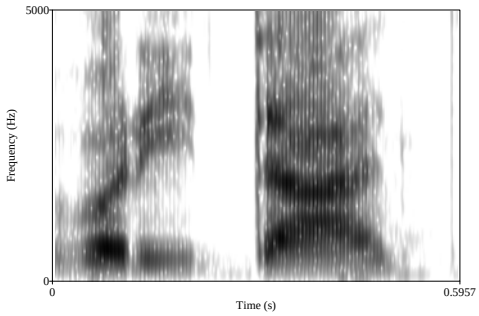

Fig. 1
Espectrograma correspondiente a la secuencia [b↔Pitát].

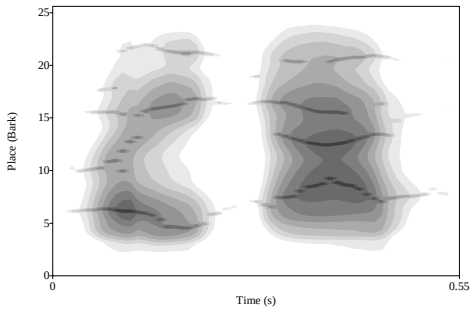

Fig. 2
Cocleagrama correspondiente a la secuencia [b↔Pitát].

En [10], reproducido a continuación, se describe muy sumariamente el proceso de audición:

> Sound waves impinge upon the outer ear, and travel down the ear canal to the eardrum. The eardrum is a thin membrane of skin which is stretched like the head of a drum at the end of the ear canal. Like the membrane of a microphone, the eardrum moves in response to air pressure fluctuations.
>
> These movements are conducted by a chain of three tiny bones in the middle ear to the fluid-filled inner ear. There is a membrane (the basilar membrane) that runs down the middle of the conch-shaped inner ear (the cochlea). This membrane is thicker at one end that the other. The thin end, which is closest to the bone chain, responds to high-frequency components in the acoustic signal, while the thick end responds to low-frequency components. Each auditory nerve fiber innervates a particular section of the basilar membrane, and thus carries information about a specific frequency component in the acoustic signal. In this way, the inner ear performs a kind of Fourier analysis of the acoustic signal, breaking it down into separate frequency components.

Como la cóclea tiene diferente grosor en sus distintas partes (figura 3), presenta una respuesta diferente para cada rango de frecuencias. Cada fibra nerviosa auditiva tiene una frecuencia característica a la que es sensible. Así, una fibra con una frecuencia característica de 2000 Hz va a responder a las vibraciones de esta frecuencia. Por eso, la escala de frecuencias en Hz no es adecuada para representar la audición humana, porque la escala en Hz es lineal, mientras que la escala auditiva no. Para eso, los Bark son unidades que reflejan mejor la percepción auditiva. Las figuras 3 y 4 muestran la relación entre ambas escalas.

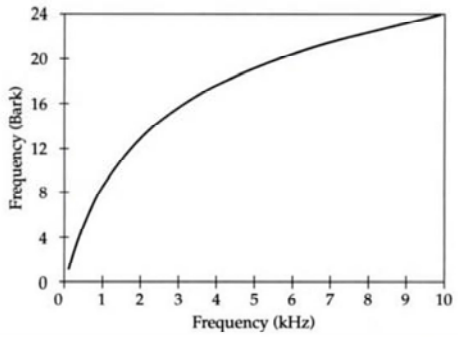

Fig. 3

Relación entre la escala de frecuencia auditiva (en Bark) y acústica (en kHz) [10].

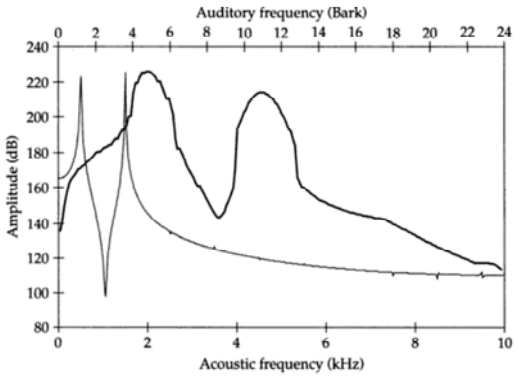

Fig. 4

Comparación de la representación acústica (línea delgada) y auditiva (línea gruesa) de una onda compleja formada por dos ondas simples de 500 y 1500 Hz [10].

3. METODOLOGÍA

Los cocleagramas no han sido tradicionalmente utilizados por los laboratorios de lingüística forense para realizar comparaciones de muestras dubitadas e indubitadas, quizás porque se han visto como gráficos que simplifican la información acústica, con lo que se pierden numerosos matices que sí es posible analizar con los espectrogramas o con otras técnicas de análisis multidimensional. Sin embargo, con el Praat [11] es posible calcular el coeficiente de correlación (r de Pearson) dados dos cocleagramas de igual duración. Este índice estadístico permite medir la relación lineal entre dos variables cuantitativas, independientemente de su escala. Un valor igual a 1 significa que los datos domparados son idénticos; cuanto menor es el valor, mayores son las diferencias observadas. Para ello voy a utilizar un script (una pequeña aplicación utilizada para automatizar tareas) para el programa Praat desarrollado por Paul Boersma (profesor de la Universidad de Ámsterdam y creador del Praat) y puesto a disposición en la lista de distribución de los usuarios de dicho programa. El script lo reproduzco en el Anexo.

Las comparaciones se han realizado con 2 grupos de datos de distinta naturaleza de 5 hablantes:

i. Sonidos dubitativos (correspondientes a pausas llenas) con el mismo timbre vocálico ([i:]) y de idéntica duración (40 milisegundos)

ii. Palabras enteras (*veritat*). Un requisito del script para calcular el coeficiente de correlación entre dos sonidos es que ambos archivos tengan exactamente la misma duración. Por ello, se ha escalado la duración a 55 milisegundos.

Para todos los ejemplos se ha normalizado la amplitud media a 70 dB con el fin de que el cálculo de la similitud entre los sonidos no se viera afectado por la intensidad de los distintos fragmentos analizados.

La tabla 2 muestra el número de fragmentos de la comparación.

Tabla 2

Resumen de los fragmentos analizados por tipo y hablante.

Hablante	**Sonidos dubitativos [i:]**	***veritat***
IB	10	10
AL	4	10
AG	11	10
MP	4	10
MC	12	10
Total	41	50

4. RESULTADOS

Se ha calculado el coeficiente de correlación (r de Pearson) para comparar la similitud entre todos los cocleagramas mediante el script de Boersma (en el anexo) para el programa Praat.

En primer lugar, se ha realizado la comparación de los sonidos dubitativos [i:], con un total de 820 comparaciones. La tabla 3 muestra las medias de los coeficientes de correlación en las comparaciones de las distintas muestras para cada hablante o pareja de hablantes. La tabla 4, por su lado, muestra las medias en las comparaciones intra-hablante e inter-hablante.

Tabla 3

Resultados del coeficiente de correlación (r de Pearson).

Hablantes	r de Pearson
AG	0,7416
AL	0,7363
IB	0,7438
MC	0,7305
MP	0,7667
AG*AL	0,7409
AG*IB	0,7143
AG*MC	0,65
AG*MP	0,5084
AL*IB	0,7433
AL*MC	0,7269
AL*MP	0,5744
IB*MC	0,7018
IB*MP	0,5785
MC*MP	0,6367

Tabla 4.

Resultados del coeficiente de correlación (r de Pearson).

	r de Pearson
Media mismo hablante (variación intra-hablante)	0,7387
Media hablantes distintos (variación inter-hablante)	0,6714

La figura 5 muestra los resultados de las comparaciones. Las cinco primeras barras corresponden a la media de los coeficientes de correlación de las comparaciones de los distintos fragmentos de cada hablante (y muestra, pues, la media de similitud de las comparaciones de realizaciones distintas de cada uno de los 5 hablantes del estudio); las demás barras muestran el grado de similitud cuando se comparan las realizaciones de hablantes distintos.

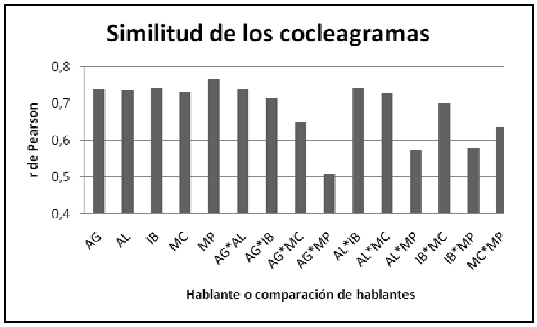

Fig. 5

Media del coeficiente de correlación de Pearson en la comparación de fragmentos de un hablante o pareja de hablantes.

La figura 6, por su lado, muestra los datos de un modo global. La barra de la izquierda indica la media de similitud cuando se comparan realizaciones de los mismos hablantes (variación intra-hablante); la barra de la derecha indica la media de similitud en la comparación de cocleagramas correspondientes a hablantes distintos.

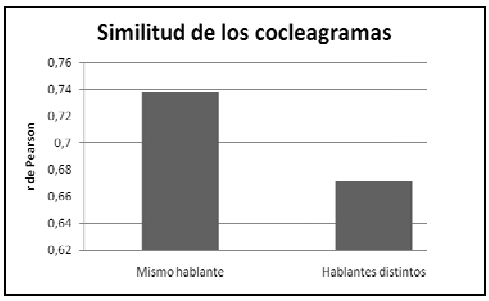

Fig. 6

Media del coeficiente de correlación de Pearson para distintas realizaciones de un mismo hablante o de hablantes distintos.

Los resultados indican, por un lado, que la variación intra-hablante es menor (y por tanto, el grado de similitud que indica el coeficiente de correlación de Pearson es mayor) que la variación inter-hablante (con un grado de similitud menor).

Los datos estadísticos confirman esta impresión: se ha realizado una prueba ANOVA para comparar los resultados de las comparaciones (r de Pearson), cuyos resultados son altamente significativos ($p<0.001$). Sin embargo, el examen atento de la figura 5 muestra que en algunas comparaciones entre hablantes distintos (como es el caso de AG*AL, AG*IB, AL*IB y AL*MC) los resultados son similares que en las comparaciones de distintas realizaciones de un mismo hablante.

En segundo lugar, presentamos los resultados de las comparaciones de la palabra *veritat*. En este análisis se han realizado 1250 comparaciones. Los resultados son similares a los de los segmentos [i:]. Las tablas 5 y 6 muestran los resultados numéricos.

Tabla 5

Resultados del coeficiente de correlación (r de Pearson).

Hablantes	**r de Pearson**
AG	0,8013
AL	0,7478
IB	0,7829
MC	0,8075
MP	0,7758
AG*AL	0,7739
AG*IB	0,7331
AG*MC	0,7856
AG*MP	0,7498
AL*IB	0,7561
AL*MC	0,8058
AL*MP	0,7531
IB*MC	0,7536
IB*MP	0,7527
MC*MP	0,7528

Tabla 6

Resultados del coeficiente de correlación (r de Pearson).

	r de Pearson
Media mismo hablante (variación intra-hablante)	0,8007
Media hablantes distintos (variación inter-hablante)	0,7601

Las figuras 7 y 8 muestran los resultados promediados de los coeficientes de correlación de Pearson para cada hablante y pareja de hablantes, por un lado, y la media de los coeficientes para las comparaciones de fragmentos pertenecientes a los mismos hablantes y a hablantes distintos, por el otro.

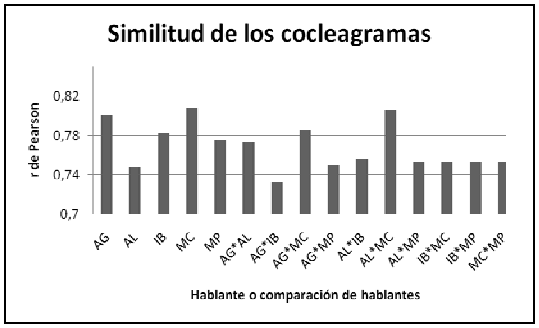

Fig. 7

Media del coeficiente de correlación de Pearson en la comparación de fragmentos de un hablante o pareja de hablantes.

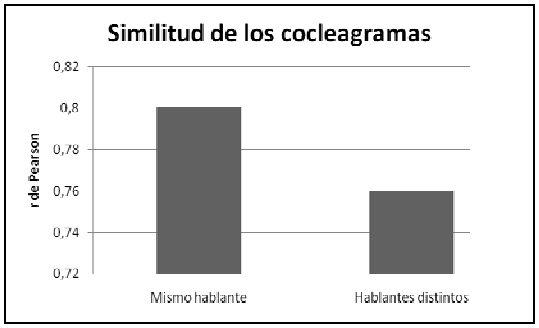

Fig. 8

Media del coeficiente de correlación de Pearson para distintas realizaciones de un mismo hablante o de hablantes distintos.

Los resultados para este conjunto de comparaciones también muestran diferencias estadísticamente significativas ($p<0.001$ en la prueba ANOVA) entre las comparaciones intra-hablante e inter-hablante. Sin embargo, en la figura 7 se observa que los resultados de las comparaciones intra-hablante para el hablante AL se obtienen resultados menores que en algunas comparaciones inter-hablantes. Aun así, globalmente se observan mayores diferencias en la comparación entre muestras de hablantes distintos que entre muestras del mismo hablante.

5. CONCLUSIONES

Los resultados de este estudio han confirmado, por un lado, que la comparación de cocleagramas puede ser útil como parámetro complementario para identificar hablantes en fonética forense. Se ha demostrado que la variación intra-hablante es, en general, menor que la variación inter-hablante, y que las diferencias son estadísticamente significativas. Sin embargo, esta técnica debe de considerarse con mucha cautela por las siguientes razones.

En primer lugar, se ha podido constatar que los resultados son desiguales. Tomando los datos globalmente, en efecto pueden hallarse diferencias significativas; sin embargo, al analizar más detalladamente cada comparación, se observa que en algunos casos comparaciones inter-hablantes muestran mayor similitud que comparaciones intra-hablantes.

En segundo lugar, se observa un mayor poder discriminante en la consideración del sonido dubitativo [i:] que en la producción [b↔Pitát]. En el primer caso, los valores de correlación en las comparaciones intra-hablante son altos y estables, mientras que en el segundo son mucho más desiguales. Además, la diferencia entre la media de los valores de correlación intra-hablante e inter-hablante es igualmente superior en el caso de los sonidos dubitativos que en el de las palabras sueltas.

Finalmente, hay que destacar que existen límites en cualquier técnica para identificar hablantes. Las identificaciones expertas de hablantes no pueden basarse únicamente en la percepción auditiva ni en técnicas que la emulen (tales como los cocleagramas), debido a los pobres resultados reportados en la numerosa bibliografía (como las referencias [3] a [9] de este artículo), por un lado, y a la transformación del señal acústico en el oído interno, por el otro. En este estudio se ha demostrado que algunas comparaciones entre muestras de hablantes distintos muestran un alto grado de similitud desde un punto de vista auditivo. Sin embargo, en el plano acústico (formantes, F0, jitter, shimmer, razón ruido-armónicos, etc.) sí es posible discriminar los hablantes en fragmentos dubitativos [12].

Por lo tanto, debe considerarse la comparación de cocleagramas como un nuevo parámetro a añadir a la cesta de parámetros útiles para la identificación forense de locutores (ver, por ejemplo, [13], entre otros), aunque en ningún caso puede ser determinante.

REFERENCIAS

[1] Nolan, F. “Auditory and acoustic analysis in speaker recognition”. En Gibbons, J. (ed.), Language and the law. London/New York: Longman. 326-345, 1994.

[2] Cambier-Langeveld, Tina. “Current methods in forensic speaker identification: Results of a collaborative exercise”. *The International Journal of Speech, Language and the Law*, 14(2), 2007.

[3] McGehee, F. "The reliability of the identification of the human voice." *Journal of General Psychology* 17: 249-271, 1937.

[4] McGehee, F. "An experimental study of voice recognition." *Journal of General Psychology* 31: 53-65, 1944.

[5] Künzel, H. J. “On tyhe problem of speaker identification by victims and witnesses”. *Forensic Linguistics*, 1: 45-58, 1994.

[6] Hollien, H. y Schwartz, R. "Aural-perceptual speaker identification: problems with noncontemporary samples". *Forensic Linguistics* 7.2, p.199-211, 2000.

[7] Foulkes, P. y Barron, A. “Telephone speaker recognition amongst members of close social network”. *Forensic Linguistics: The International Journal of Speech, Language and the Law* 7, 180-198, 2000.

[8] Blatchford, H. y Foulkes, P. “Identification of voices in shouting”. *The International Journal of Speech, Language and the Law* 13, 241-254, 2006.

[9] Schiller, N. O. y Köster, O. "The ability of expert witnesses to identify voices: A comparison between trained and untrained Listeners". *Forensic Linguistics* 5: 1–9, 1998.
[10] Johnson, K. (2a edición). *Acoustic and Auditory Phonetics*. Blackwell Publishing, 2003.
[11] Boersma, P. y Weenink, D. *Praat: doing phonetics by computer* (Version 4.5.08) [Programa informático]. Disponible en http://www.praat.org/.
[12] Cicres, J. "Análisis discriminante de un conjunto de parámetros fonético-acústicos de las pausas llenas para identificar hablantes". *Síntesis Tecnológica* 3 (2), p. 87-96, 2007.
[13] Rose, Ph., *Forensic Speaker Identification*, Taylor & Francis, Londres, 2002.

ANEXO: Script de Paul Boersma para la comparación de cocleagramas

```
if numberOfSelected ("Cochleagram") <> 2
exit Please select two cochleagrams first.
endif
To Matrix
cochleagram1 = selected ("Matrix", 1)
cochleagram2 = selected ("Matrix", 2)
select 'cochleagram1'
Rename... coch1
nt1 = Get number of columns
nf1 = Get number of rows
select 'cochleagram2'
Rename... coch2
nt2 = Get number of columns
nf2 = Get number of rows
if nt1 <> nt2 or nf1 <> nf2
exit The two cochleagrams should have the same size.
endif
Create Table... table nt1*nf1 2
Set column label (index)... 1 Coch1
Set column label (index)... 2 Coch2
Formula... Coch1 Matrix_coch1 [(row-1) div nt1 + 1, (row-1) mod nt1 + 1]
Formula... Coch2 Matrix_coch2 [(row-1) div nt1 + 1, (row-1) mod nt1 + 1]
Report correlation (Pearson r)... Coch1 Coch2 0.025
```