PERCEPCIÓN AUDIOVISUAL DE LAS VOCALES DEL ESPAÑOL EN CONDICIONES UNIMODALES Y BIMODALES CONGRUENTES E INCONGRUENTES

AUDIOVISUAL PERCEPTION OF CONGRUENT AND INCONGRUENT SPANISH VOWELS IN UNIMODAL AND BIMODAL CONDITIONS

ENRIQUE SANTAMARÍA BUSTO
New York University in Madrid
enrique.santamaria@nyu.edu

Artículo recibido el día: 16/01/2015
Artículo aceptado definitivamente el día: 04/05/2016
Estudios de Fonética Experimental, ISSN 1575-5533, XXV, 2016, pp. 81-148

RESUMEN

El presente trabajo propone una primera aproximación al estudio de la integración de la señal visual y auditiva en los procesos de percepción bimodal de las vocales del español. Siguiendo el experimento de McGurk y MacDonald (1976), se pretende averiguar si las claves discordantes en la percepción audiovisual pueden alterar la identificación del estímulo auditivo, y si esto provoca un resultado perceptivo distinto al que se produce en cada canal por separado. Estos resultados nos permitirán saber el grado de influencia de la señal visual sobre la señal sonora para esta clase de estímulos y condiciones, y valorar por otro lado si los resultados permiten apuntar diferencias de sensibilidad hacia la señal visual en función del sexo del hablante. Con este objetivo, 28 sujetos (12 hombres y 16 mujeres) tuvieron que identificar las 5 vocales del español distribuidas en tres bloques: a) 25 combinaciones audiovisuales cruzadas entre todas las vocales (por tanto, combinadas de forma congruente e incongruente); b) 10 estímulos visuales distribuidos en dos series (5 x 2); y c) 5 estímulos auditivos. Junto a ello, los participantes tuvieron que señalar el grado de seguridad con el que emitían su respuesta. Los resultados reflejan, entre otros datos, que la presencia de estímulos visuales incongruentes afecta la percepción de la señal auditiva y se producen casos de fusiones; que la información visual por sí sola no es suficiente para la discriminación, aunque se aprecian diferencias entre vocales; y que se encuentran diferencias significativas en la percepción visual y la variabilidad de las respuestas entre hombres y mujeres.

Palabras clave: *percepción audiovisual, vocales, efecto McGurk, lectura labiofacial.*

ABSTRACT

This paper proposes a preliminary approach to the study of the integration of auditory and visual signals in the process of audiovisual speech perception of Spanish vowels. Following the experiment of McGurk and Macdonald (1976), the purpose of this study is to find out whether incongruent cues in audiovisual speech perception can alter the identification of the auditory stimulus, and to analyze whether this bimodal integration process causes a perceptual result other than the one that occurs in each channel separately. These results will allow us to know the extent to which the visual signal affects this class of stimuli and conditions, and determine, on the other hand, if the findings reflect differences in sensibilities in regards to the visual channel depending upon the speaker's gender. For this purpose, 28 subjects (12 men and 16 women) had to identify the five vowels of

Spanish distributed at random into three blocks: a) 25 audiovisual cross combinations between all vowels (thus in congruent and incongruent audiovisual conditions); b) 10 visual stimuli distributed in two series (5 x 2); and c) 5 auditory stimuli. Along with this, the participants had to indicate the degree of confidence of their response. Results showed, among other findings, that the presence in this kind of incongruent visual stimuli affects the perception of the auditory signal and results in perceptual fusions. It also demonstrates that visual information alone is not sufficient for discrimination, although important differences can be found depending on the presence of visual cues, such as rounding or vowel openness. There are also significant differences in visual perception and the variability of responses between men and women.

Keywords: *audiovisual speech perception, vowels, McGurk effect, lipreading.*

1. INTRODUCCIÓN

La enorme variedad de estímulos que recibe el ser humano convierte al cerebro en un complejo procesador perceptivo capaz de recoger e interpretar señales sensoriales distintas y asincrónicas pertenecientes a un mismo evento. De todas las señales, la información multisensorial obtenida a través de la imagen y del sonido ha sido una de las más analizadas. Estímulos como los de un aplauso, o el movimiento de un cohete con su correspondiente explosión se reconstruyen simultáneamente en el cerebro, a pesar de que por razones físicas (la luz viaja a mayor velocidad que el sonido) y fisiológicas (el tiempo de las señales visual y auditiva en llegar a la corteza cerebral primaria es distinto), se produce un desajuste temporal entre ellos (Velasco *et al*, 2011). El cerebro refleja por tanto un comportamiento adaptativo de suma importancia, ya que integra y reajusta los estímulos recibidos para ofrecer una reconstrucción coherente del mundo que experimentamos (O'Shea, 2005).

Uno de los procesos más estudiados de integración multisensorial entre imagen y sonido se produce en el habla. Durante años se ha visto cómo la observación del rostro y los movimientos articulatorios del hablante se complementa con la escucha, y ayuda a la percepción del acto verbal comunicativo. Para activar este proceso los hablantes realizan una lectura labiofacial del interlocutor que se superpone a la percepción auditiva. Esta percepción bimodal opera de modo inconsciente, y solamente parecemos darnos cuenta de su procesamiento cuando nos enfrentamos a situaciones a las que no estamos acostumbrados. Algunos

ejemplos recogidos por Rahmawati y Ohgishi (2011) o Massaro (1998), señalan en este sentido observaciones como la preferencia de muchos hablantes por mantener conversaciones físicas y no por teléfono; la irritación que producen las películas extranjeras mal dobladas (es decir, con mala sincronización de voz e imagen); los comentarios de personas que señalan «oír» mejor la televisión con las gafas puestas; la molestia que supone el desfase temporal entre las señales visual y auditiva en las comunicaciones audiovisuales por Internet (videollamadas); o los estudios que muestran avances más lentos en la adquisición del habla en niños ciegos frente a niños sin problemas visuales.

Durante mucho tiempo se pensó que la lectura labiofacial era un recurso usado fundamentalmente por personas sordas, aunque a partir de 1950 diversos estudios demostraron cómo, en ambientes ruidosos en los que la relación señal-ruido es desfavorable para la señal, la visibilidad del rostro del interlocutor puede mejorar significativamente la inteligibilidad del habla (Sumby y Pollack, 1954; Erber, 1969; MacLeod y Summerfield, 1987; Ross *et al*, 2007). La percepción del habla se consideró entonces durante cierto tiempo como un proceso de naturaleza fundamentalmente unimodal, de carácter auditivo, en el que solo en casos de personas con problemas auditivos o en situaciones extraordinarias como en los ambientes con ruido intervenía la vista.

Los estudios sobre la percepción audiovisual del habla dieron sin embargo un vuelco en 1976, con la publicación de un influyente artículo de McGurk y MacDonald titulado *Hearing lips and seeing voices*. En este estudio se mostraba cómo estímulos auditivos basados en la repetición de secuencias CV a partir de segmentos oclusivos y vocal abierta [baba], [gaga], [papa] y [kaka] eran percibidos en su gran mayoría de forma distinta si al estímulo auditivo se le superponía un vídeo incongruente del hablante pronunciando otra de esas secuencias. En su experimento, el estímulo auditivo [baba] combinado con la imagen del hablante articulando [gaga], era percibido por la mayor parte de los hablantes como [dada], mientras que el auditivo [papa] combinado con el visual [kaka] era percibido como [tata]. El resultado mostraba por tanto una destacada presencia de una sílaba nueva que no existía en las señales auditiva y visual por separado, dando lugar bien a una fusión de los dos estímulos, o bien, incluso, a una combinación de ambos, como en los ejemplos en los que la secuencia auditiva [gaga] combinada con la imagen de [baba] era percibida como [gabga], [bagba], [baga] o [gaba]. Por otra parte el experimento mostraba también casos, más minoritarios, en los que el estímulo percibido procedía de la señal visual, y no de la auditiva, como se hubiera esperado. Estos efectos se revelaron por tanto extraordinarios, ya que los oyentes, influidos por la señal visual, percibían mayoritariamente un estímulo que era

distinto de la señal auditiva, incluso cuando esta señal lograba identificarse de forma aislada sin ninguna dificultad.

Unos resultados similares a los de McGurk y McDonald se obtuvieron también en experimentos realizados al año siguiente, como los de Dodd (1977), o Yonovitz *et al*. (1977), donde en secuencias CV en las que se variaba la consonante en condiciones de percepción bimodal incongruente, los jueces señalaban consonantes que en determinadas combinaciones no existían en ninguna de las dos condiciones. La característica de las consonantes percibidas es que estas solían mantener el modo de articulación y la sonoridad de la consonante presentada auditivamente, pero no tanto el lugar de articulación, más influenciado por la señal visual. Se demostraba por tanto que aunque la capacidad de obtener información variaba entre los sujetos y los estímulos presentados, el procesamiento se realizaba siempre a partir de la información suministrada por las dos señales. Por otro lado, cuando los oyentes cerraban los ojos y escuchaban los estímulos sin mediación visual, no se producía ninguna de estas confusiones.

La existencia de este fenómeno (conocido en adelante como «efecto McGurk») demostraba con ello que la percepción visual de los movimientos articulatorios del hablante se activaba por el interlocutor en todas las condiciones de comunicación audiovisuales, y no solo en circunstancias excepcionales, lo que probaba que la percepción del habla no es únicamente un procesamiento perceptivo de información auditiva. A su vez, se lograba probar también que la experiencia de percepción visual podía modificar significativamente la experiencia de la escucha, lo que sugería que los oyentes extraían y utilizaban las informaciones que cada canal le proporcionaba antes de tomar una decisión perceptiva.

Tras la publicación de este artículo, el interés por entender mejor el procesamiento de percepción del habla bimodal dio lugar a numerosas investigaciones que modificaron y examinaron distintas variables en la presentación de los estímulos. Se comprobó así que la integración audiovisual se producía no solamente a nivel de sílabas o de palabras (Ostrand *et al*, 2011), sino también de frases tomadas del habla real que se entendían mejor a nivel audiovisual que auditivo, mostrándose especialmente relevante en ambientes con ruido (MacLeod y Summerfield, 1987). También se vio que cuando se alteraban los parámetros suprasegmentales por parte de hablantes con acento extranjero, o cuando el texto era cognitivamente complejo, la comprensión por parte de hablantes nativos resultaba más fácil si los estímulos eran audiovisuales que si eran solamente auditivos (Reisberg *et al,* 1987; Kim y Davis, 2003). La integración audiovisual también pudo compararse desde una perspectiva neurofisiológica a través de estudios que demostraban que la

información visual obtenida de la lectura labiofacial activaba diversas zonas del cerebro, como las áreas corticales auditivas (Sams *et al*, 1991; Calvert *et al*, 1997), las áreas cerebrales motoras encargadas de ordenar la articulación de sonidos (Skipper *et al*, 2007), o las neuronas llamadas «espejo» que operan en el sistema motor (Rizzolatti y Craighero, 2004).

Por su parte, el efecto McGurk se caracterizó a partir de entonces como un fenómeno recurrente que aparecía en los estudios de percepción del habla bimodal en situaciones de incongruencia entre la señal visual y auditiva. La ilusión del efecto McGurk es tan robusta que aparece incluso en situaciones en las que el sujeto conoce el fenómeno (Summerfield y McGrath, 1984); cuando existe una asincronía entre estímulos (van Wassenhove, 2007) o se introducen cambios de *tempo* en las señales (Munhall *et al*, 1996); cuando la señal acústica es degradada por un filtro de baja frecuencia o se elimina información espectral (Green y Norrix, 1997); cuando se suprime parcialmente alguno de los estímulos (Munhall y Tokhura, 1998); cuando las señales auditiva y visual proceden de hablantes de distinto sexo (Green *et al*, 1991) o de distinto origen cultural (Rahmawati y Ohgishi, 2011); cuando la señal visual es una animación (Massaro y Cohen, 1990); cuando se invierte la posición del rostro (Campbell, 1994; Massaro y Cohen, 1996) o el brillo de la imagen (Kanzaki y Campbell, 1999); o incluso cuando el rostro del hablante se sustituye por puntos luminosos situados en los articuladores (Rosenblum y Saldaña, 1996), o se distorsiona mediante el uso de filtros que «pixelan» la imagen (MacDonald *et al*, 2000).

A su vez, se observó también que el efecto McGurk no se produce únicamente en procesos perceptivos de integración audiovisual con claves discordantes, sino que también aparece con estímulos sensosomáticos en los que un soplo de aire que simula la aspiración en el momento de la emisión puede provocar la percepción de /pa/ en lugar de /ta/ (Gick y Derrick, 2009). Otros estudios demuestran cómo puede producirse el mismo efecto cuando el estímulo visual procedente de los movimientos articulatorios del hablante se sustituye en este caso por el tacto de su boca y rostro (Fowler y Dekle, 1991). Esto demuestra que la percepción del habla está sujeta a factores multisensoriales que pueden ir más allá de la bimodalidad audiovisual, afectando en su proceso de reconocimiento e integración la percepción aislada de la señal sonora.

El análisis de los datos obtenidos por el efecto McGurk dio lugar a varias conclusiones. En primer lugar, se reconoce que el habla unimodal de tipo auditivo transmite más información que la de tipo visual (Massaro, 1998), y que es más fácil identificar sílabas a partir de solo una señal de audio que solo de vídeo, ya

que en el habla, una gran parte de lo que ocurre en el tracto vocal es invisible (Girin *et al*, 2001). Palabras como *bato*, *pato* y *mato*, o sílabas como *aga* o *aka* no pueden distinguirse visualmente si no es con ayuda auditiva[1]. El gran reto que tanto la percepción bimodal como el efecto McGurk suponen para las teorías de percepción del habla consiste precisamente en saber por qué se produce una integración audiovisual en el procesamiento perceptivo, cuando la información que llega por el canal auditivo es por sí sola suficiente en la mayor parte de los contextos comunicativos desarrollados entre hablantes que se comunican en ambientes sin ruido y no tienen problemas auditivos. Esto ha alimentado en las últimas décadas la publicación de numerosos estudios que intentan dilucidar si la percepción del habla es en esencia de naturaleza auditiva o visual, dando lugar a los distintos postulados teóricos surgidos en Fonética Perceptiva: Teoría Articulatoria o Motora; Análisis por Síntesis; Realista-Directa; Modelo de Percepción de Lógica Difusa; Teoría Auditiva, Modelos Cognitivos, etc.

Con la publicación del efecto McGurk, algunos autores como los propios McGurk y MacDonald (1976), Walden *et al.* (1977), MacLeod y Summerfield (1987) o Munhall *et al.* (1996), interpretaron que el estímulo visual juega un papel importante para informar sobre el lugar de articulación de algunas consonantes, mientras que el modo de articulación (oral / nasal; sordo / sonoro) podía percibirse más claramente a partir de la información auditiva. A partir de esto, autores como Robert-Ribes *et al.* (1998) señalaron la complementariedad de las señales visual y auditiva en un primer nivel informativo que recoge las claves enviadas por cada canal antes de ser procesadas en el cerebro. Esto explicaría por qué en la percepción bimodal las claves perceptivas de los estímulos se transmiten mejor a través de la vista y del oído que de forma individual. Otros autores como Massaro y Cohen (1990), o el mismo Massaro (1998), matizaron también las conclusiones sobre la especialización que, a priori, pudiera atribuirse a cada uno de los canales, señalando que la señal auditiva, por sí sola, era también muy importante para percibir el lugar de articulación, por lo que no podía hablarse *stricto sensu* de sensibilidades distintas de cada canal perceptivo como instrumento para interpretar las claves que intervienen en la percepción del habla. Las informaciones transmitidas por tanto a través de estos canales son complementarias, pero no específicas ni excluyentes en cada canal.

[1] En un experimento de lectura labiofacial en el que se analizaron nueve grupos de fonemas que son visualmente indistinguibles (llamados visemas), Walden *et al.* (1977) encontraron que si dos consonantes pueden confundirse entre ellas y su distinción es la sonoridad, se prefiere el segmento sordo (excepto para /p/ vs /b/); y si la distinción es la nasalidad, el segmento oral.

El grado de relevancia que cada uno de estos canales tiene en la percepción del habla bimodal no está sujeto tampoco al consenso entre investigadores, y varía siempre según las variables. Así, en el estudio seminal de McGurk y MacDonald, la información visual tiene para sus autores un efecto de suma importancia, y opera con tanta fuerza como la información auditiva, como puede verse en los procesos de fusión y combinación de los estímulos percibidos cuando ambas señales se presentan de forma incongruente. Un caso semejante parece ocurrir en estudios de prosodia audiovisual en los que la imagen suele tener un peso tan importante o incluso mayor que la información auditiva, aunque esta relación es diferente según las emociones (Abelin, 2007).

Estos resultados arrojan por tanto ciertas matizaciones. En este sentido, una de las teorías más extendidas entre los estudios que analizan los procesos de percepción bimodal del habla, el Modelo de Percepción de Lógica Difusa (Massaro, 1989), sostiene que una modalidad no domina a la otra, sino que ambas se integran en un proceso en el que los receptores evalúan las informaciones, y la señal más consistente y menos ambigua es la que tiene mayor influencia (Massaro, 1998). Esta teoría aparece avalada por el hecho de que el efecto McGurk puede modularse en función del peso que se dé a cualquiera de los estímulos audiovisuales discordantes (Colin *et al*, 2005). De este modo la influencia de la información visual resulta más relevante cuando la señal auditiva tiene menor intensidad, se le añade ruido, se modifica su inteligibilidad, o sufre un intervalo de demora (Munhall *et al*, 1996), mientras que la señal auditiva resulta más influyente cuando la señal visual se hace borrosa, es rotada, invertida, o se minimiza de tamaño (MacDonald, 2006).

La mayor parte de los estudios que ha investigado el efecto McGurk se ha centrado tradicionalmente en la percepción de consonantes insertadas dentro de secuencias silábicas del tipo CV. La razón principal en la elección de estos segmentos estriba en la destacada información visual que pueden ofrecer algunos de ellos a partir de su lugar de articulación (Green y Kuhl, 1989). Así, en su estudio sobre la contribución visual de las consonantes a la inteligibilidad del habla, Nielsen (2004) demostró que la mayor parte de los segmentos analizados (15 sonidos consonánticos del inglés insertados en palabras reales) ayudaban y mejoraban la comprensión del habla en condiciones audiovisuales con diferentes niveles de ruido, especialmente los sonidos interdentales y labiodentales[2].

[2] Sin embargo también había otros, como [r], que suponían un deterioro de la comprensión en esas mismas condiciones. En este caso los interlocutores tendían a confundirlo con [w].

Teniendo en cuenta la potencialidad de información visual que ofrecen algunos segmentos consonánticos por su lugar de articulación, el estudio de las vocales ha sido más minoritario. Sin embargo, experimentos como los de Summerfield y McGrath (1984) o Green y Gerdeman (1995) para el inglés; Robert-Ribes *et al.* (1998) o Lisker y Rossi (1992) para el francés; Traunmüller y Öhrström (2007) para el sueco; o el de Valkenier *et al.* (2012) para el neerlandés, demostraron que la percepción de vocales en condiciones audiovisuales incongruentes puede producir también en distintas lenguas decisiones perceptivas que arrojan luz sobre el peso que tiene cada señal en el proceso de percepción bimodal, tal y como ha podido comprobarse a su vez desde un punto de vista neurofuncional, que localiza específicamente el proceso perceptivo para estos estímulos en el surco temporal medio superior (Murase *et al*, 2008). La hipótesis sobre la que se diseñaron estos experimentos sobre vocales consideraba que en los procesos perceptivos de integración audiovisual, la información visual contribuye con fuerza a la percepción vocálica, ya que las vocales poseen determinados gestos articulatorios propios que resultan visibles y relevantes para su identificación, tales como la acción de los labios y en menor medida la altura de la lengua, cuyo acercamiento y alejamiento con respecto del paladar origina distintos grados de abertura.

Tradicionalmente las vocales se describen desde un punto de vista articulatorio a partir de tres dimensiones: las dos primeras hacen referencia a la posición de la lengua (altura y avance o retroceso del cuerpo lingual), mientras que la tercera hace referencia a la disposición de los labios o redondeamiento labial. Aunque las vocales de una misma lengua pueden variar en cualquiera de estas tres dimensiones[3] de un dialecto a otro[4] (Ladefoged, 2001), el redondeamiento puede

[3] Aunque la Asociación de Fonética Internacional describe, a partir de los parámetros de Jones (1917), un esquema de solo dos posiciones a nivel de la acción labial (redondeado / no redondeado), frente a las cuatro posiciones cardinales en la dimensión de altura o las tres posiciones en el eje antero-posterior, es evidente que el redondeamiento vocálico no es una propiedad en sí, sino que esta difiere también en grados, relevantes desde un punto de vista fonético, y en su mayoría relacionados con la abertura y la posición de la lengua. Así, como señalan Ladefoged y Maddieson (1996:292-297), por un lado la altura del cuerpo lingual y el redondeamiento labial suelen guardar una estrecha relación, de modo que cuanto más cerrada es una vocal más acusada es su labialización, aunque hay excepciones; por otro lado la mayoría de las lenguas del mundo presentan una relación entre anterioridad y no redondeado y posterioridad-redondeado, aunque por supuesto pueden encontrarse lenguas con vocales anteriores redondeadas y posteriores no redondeadas. Por otra parte existen también lenguas que presentan hasta tres tipos contrastivos de redondeamiento vocálico, tal como ocurre en sueco o algunos dialectos del noruego, que distinguen tres vocales altas anteriores diferenciadas por el grado de redondeamiento. Considerándolo todo, aunque desde un punto de vista fonológico una clasificación binaria resulta suficiente para la

ser en algunas lenguas una cualidad distintiva independiente que por sí sola puede diferenciar entre pares de vocales. Esto ha hecho que la bibliografía que tradicionalmente ha estudiado estos tipos de percepción audiovisual se haya centrado en el análisis de lenguas cuyos inventarios vocálicos presentaran vocales opuestas únicamente a partir del rasgo [+/-redondeado], como las vocales anteriores [i / y, I / Y, e / ø, e / Y, ɛ / œ], tal y como ocurre con algunas vocales del francés y la mayor parte de las lenguas germánicas y fino-ugrias, entre otras, analizadas en estos trabajos.

El estudio de estas oposiciones vocálicas en condiciones unimodales y bimodales congruentes e incongruentes permite saber por un lado si la identificación del redondeamiento vocálico, que es, de las tres dimensiones, la que más información visual ofrece para identificar a una vocal (Tseva, 1989, *apud* Lisker y Rossi, 1992: 394), se hace mejor a través de la información que llega por el canal visual (es decir, mediante lectura labiofacial), o si lo hace mejor de forma auditiva. Por otro lado, el estudio permite saber también si en la percepción vocálica se producen fusiones análogas a las descritas por McGurk y MacDonald (1976), y de ser así, cómo se producen estas fusiones en función de la información que llega por cada canal.

Los resultados de algunos de estos estudios (Traunmüller y Öhrström, 2007), mostraron para el sueco cómo en condiciones unimodales visuales las vocales redondeadas de los pares estudiados se percibían mejor que de forma auditiva, y que la integración que surgía a partir de condiciones bimodales incongruentes daba ejemplos de fusión en donde el estímulo percibido no se correspondía con la señal visual ni con la auditiva. Por otra parte, se vio que bajo condiciones audiovisuales el canal visual transmite mejor el redondeamiento que la abertura, de modo que aun siendo esta última una dimensión percibida también a nivel visual, es sin embargo más dependiente del canal auditivo. A esta conclusión llegaron también otras investigaciones para el francés, como las de Robert-Ribes *et al.* (1998), que señalaron que el efecto McGurk era más fuerte si la señal visual aportaba un

clasificación de la mayor parte de los inventarios vocálicos de las lenguas desde el rasgo [+/- redondeado], una descripción fonética más adecuada debería dar cuenta de los distintos tipos o grados de redondeamiento vocálico con que se producen estas vocales.

[4] En este sentido el español es una lengua que, desde un punto de vista segmental, muestra mayor variabilidad fonética entre sus diferentes dialectos a partir de las consonantes, y no tanto de las vocales, como sí ocurre en cambio en otras lenguas, como el inglés, con inventarios vocálicos más amplios (Ladefoged, 2001).

estímulo con redondeamiento labial[5]. Estos resultados apoyan también estudios que demuestran que el acceso a la información visual ayuda a mejorar la pronunciación de vocales redondeadas y no redondeadas de lenguas extranjeras que no existen en la L1 de los hablantes, particularmente en lo que tiene que ver con la dimensión anteroposterior, que a su vez se debe en parte a la influencia del parámetro [+/- redondeado] (Richardson, 2010)[6].

Otros estudios, como los de Lisker y Rossi (1992), plantearon incluso la posibilidad de que el redondeamiento labial como dimensión independiente no se percibiera auditivamente de modo fiable, de modo que la información visual, aun siendo redundante, resultara esencial para la descripción de este tipo de vocales. Para ello se diseñó un experimento en el que 20 hablantes de francés con entrenamiento fonético tenían que indicar si 18 vocales diferenciadas por su disposición labial (15 vocales propias y 3 ajenas al vocalismo francés) eran o no redondeadas, a partir de estímulos presentados en condiciones unimodales y bimodales congruentes e incongruentes. Los resultados mostraron que la habilidad para percibir auditivamente vocales redondeadas se circunscribe únicamente a las que son distintivas dentro de los inventarios fonológicos vocálicos de esos hablantes. De este modo, los hablantes de francés fueron expertos en separar auditivamente las vocales que conocían por su abertura y redondeamiento (vocales anteriores no bajas), pero alcanzaron resultados similares que otros hablantes (anglohablantes en este caso) para aquellas vocales situadas en un espacio vocálico en el que en ambas lenguas el redondeamiento no tiene una cualidad fonológica contrastiva. A nivel visual en cambio, la percepción del redondeamiento fue bastante consistente y precisa para todas las vocales presentadas, y desde el punto de vista audiovisual en condiciones incongruentes se mostró que o bien

[5] Según estos estudios, mientras que el canal visual transmite mejor el redondeamiento que las dimensiones de altura o anteroposterioridad del cuerpo lingual, el canal auditivo es mejor para transmitir la altura sobre la anteroposterioridad o la disposición labial.

[6] Como se sabe, la elevación del cuerpo lingual tiene su correlato acústico en el primer formante (F1), y el avance o retroceso en su segundo (F2), que será más agudo cuanto más anterior se articule la vocal. El redondeamiento labial tiene por su parte un correlato acústico en el tercer formante (F3) (cuanto más redondeados, más grave), pero al mismo tiempo potencia el efecto acústico en la F2, ya que tiende también a disminuir sus valores, reforzando por tanto su «distintividad» anteroposterior (Ladefoged y Johnson, 2006). De este modo, aunque el acceso de un grupo a la información audiovisual no haya mejorado significativamente en este estudio la producción [+/- redondeada] de la vocal con respecto al grupo que no tenía la información visual, puede derivarse, tal y como sostiene la autora, que la información visual sobre ese redondeamiento sí ha podido influir en cambio para mejorar el avance o retroceso del cuerpo lingual.

predominaba el estímulo que ofrecía una información menos ambigua (es decir, aquel que desde el canal visual o auditivo había sido considerado mayoritariamente como redondeado o no redondeado) o bien se producían fusiones que reflejaban ambos componentes[7].

En otros estudios, la percepción audiovisual de vocales se ha centrado en los procesos de coarticulación con consonantes (Green y Gerdeman, 1995), o comparando la percepción en contextos sin ruido y con ruido blanco (véase Valkenier *et al,* 2012, para las vocales anteriores cerradas y semicerradas /i, y, e, Y/ del neerlandés). Green y Gerdeman (1995), basándose en estudios que demuestran cómo la percepción de un segmento ocurre siempre en relación con los contextos fonéticos que lo rodean (de modo que la percepción de un sonido consonántico está influido por la vocal que lo acompaña, y del mismo modo, la identificación de una vocal comienza y se refuerza en la transición desde y hacia la consonante), plantearon un experimento de crosmodalidad audiovisual con claves congruentes e incongruentes a partir de los contextos CV propuestos por McGurk y MacDonald (1976), pero en este caso sustituyendo no solo la consonante inicial, sino también la vocal que lo acompaña, a partir de las consonantes /b, g/ y las vocales /a, i/.

El estudio, dividido en tres experimentos, permite ver respectivamente cómo: (1) las discrepancias en la cualidad vocálica entre la señal visual y auditiva se detectan fácilmente por los hablantes[8]; (2) esta discrepancia en la cualidad vocálica disminuye significativamente la magnitud del efecto McGurk en los procesos de fusión (por ejemplo /bi/ auditivo frente a /ga/ visual ofrece menos fusiones a <d> o <th> que /ba/ auditivo frente a /ga/ visual[9]); y (3) estas discrepancias en la categoría vocálica aumentan el tiempo para identificar la consonante inicial,

[7] Es preciso destacar que en los casos en los que los estímulos aportaban una información visual y auditiva muy distinta, el canal auditivo ofrecía una influencia más fuerte que el canal visual a la hora de determinar si la vocal era o no redondeada. Sin embargo este aspecto puede verse alterado por el hecho de que se pide a los hablantes que elijan la opción auditiva en caso de duda.

[8] Como se ha visto, el hecho de que los hablantes sean conscientes de las discrepancias entre la señal visual y auditiva no implica necesariamente una reducción del efecto McGurk (véanse por ejemplo Green *et al,* 1991, para estudios de crosmodalidad con hablantes de distinto sexo, o Summerfield y McGrath, 1984, en cuyo estudio algunos sujetos son conscientes del fenómeno).

[9] En el experimento los sujetos tienen que elegir entre 6 letras dadas. Los patrones de respuesta también varían entre <d> o <th> en función del contexto vocálico.

incluso cuando la información consonántica procedente de los dos canales es consistente. Los resultados de este experimento permiten ver, por un lado, cómo a partir de la percepción audiovisual vocálica el efecto McGurk no es inmune a cualquier incongruencia detectada entre los canales auditivo y visual, y por otro, que la integración de la información que llega de estos dos canales durante el proceso de percepción se realiza procesando la información coarticulatoria, de modo que el sistema perceptivo resulta sensible a discrepancias crosmodales, al menos en este tipo de contextos CV.

La relevancia que estos estudios dan al carácter bimodal en los procesos de percepción del habla, y las pruebas de que en estos procesos el alineamiento audiovisual discordante puede afectar negativamente la percepción, ponen de manifiesto la importancia que tiene la concordancia audiovisual para garantizar una exitosa integración de los dos canales. En la actualidad, las investigaciones sobre vídeos animados, asistentes virtuales, tecnologías de videollamada, o el uso de transductores en implantes cocleares o de dispositivos auditivos en personas con problemas de audición consideran estos procesos, y beben de estas fuentes (Rouger *et al*, 2008). Por otra parte, algunos estudios han demostrado que en entornos con ruido la inteligibilidad auditiva de las consonantes se ve más afectada que el de las vocales, pues estas no solo mantienen mejor los picos espectrales asociados con los formantes, sino que mucha de la información espectral situada entre estos se pierde (Assmann y Summerfield, 2004)[10]. La contribución de los sonidos vocálicos a la inteligibilidad auditiva del habla es también mayor que la de los sonidos consonánticos en personas con implantes cocleares, más sensibles a la pérdida de frecuencias altas asociadas con las consonantes (Kewley-Port *et al*, 2007). Es por tanto necesario que los estudios de percepción audiovisual, más centrados a nivel segmental en las unidades consonánticas, sigan trabajando en los efectos que la información visual vocálica discordante tiene en distintos hablantes y condiciones. La presente investigación se realiza pues en este amplio marco de estudios de crosmodalidad, con el fin de aportar cómo se comporta en primer lugar la

[10] La habilidad para identificar sonidos vocálicos sobre consonánticos en entornos con ruido aparece también incluso en los peores contextos perceptivos posibles, cuando los oyentes se enfrentan a situaciones en las que varios hablantes están hablando simultáneamente (Darwin y Carlyon, 1995). En un estudio de Scheffers (1983) sobre la identificación de vocales producidas de forma simultánea, se demostró cómo esta identificación era posible incluso cuando las vocales compartían la misma F0 (los resultados eran mejores si la diferencia era de al menos un semitono). La conclusión postulada es que el procesamiento perceptivo elabora una única vocal coloreada con algunas de las características fonéticas de la segunda vocal (por tanto una combinación de ambas orientada más hacia una de ellas), sin que exista una separación de la información espectral.

integración audiovisual en los procesos de percepción bimodal de las vocales del español en condiciones sin ruido y con hablantes sin problemas auditivos.

2. PLANTEAMIENTO DEL TRABAJO

Hasta donde este autor tiene noticia, no existe ningún estudio que analice la percepción vocálica audiovisual del español en condiciones unimodales y bimodales congruentes e incongruentes, y considere, tomando en cuenta todo su inventario vocálico, la magnitud de un posible efecto McGurk.

Siguiendo por tanto el estudio de McGurk y MacDonald (1976), el presente trabajo pretende averiguar si las claves discordantes en la percepción bimodal de las vocales del español pueden alterar la identificación del estímulo auditivo, y si esto provoca un resultado perceptivo distinto al que se produce en cada canal por separado. El estudio pretende obtener con ello nuevos datos para el español sobre la relevancia de ambos canales en casos en los que, cuando se ofrece un estímulo audiovisual discordante, los sujetos se decantan por la información ofrecida por la señal más consistente. Estos resultados nos permitirán saber el grado de influencia de la señal visual sobre la señal sonora para esta clase de estímulos y condiciones. El trabajo se centra pues en la validación de varias hipótesis relacionadas con la percepción del habla:

> (H1): La información transmitida por el canal auditivo es mayor que la del canal visual en condiciones unimodales de percepción vocálica. La percepción auditiva de las vocales del español será perfecta o casi perfecta en condiciones en las que como aquí se proponen no interviene el ruido y los hablantes no tienen problemas auditivos. La lectura labiofacial ofrecerá resultados de identificación más bajos, al ofrecer señales de identificación menos consistentes. Se espera también que en condiciones audiovisuales incongruentes la selección perceptiva en estos casos sea auditiva, y no visual.
>
> (H2): En condiciones unimodales visuales, las vocales /a, o, u/ tendrán porcentajes más altos de identificación visual que /e, i/ debido a sus claves visuales, como el mayor grado de abertura o la presencia de redondeamiento labial. Se espera por tanto que en condiciones audiovisuales incongruentes la destacada información visual que ofrecen estas vocales incida con mayor intensidad en una incorrecta identificación

de la señal auditiva, de modo que estas vocales reflejarán un mayor número de casos de identificación visual y de fusión de estímulos (un efecto McGurk). A su vez, se espera que las vocales posteriores /o, u/ de tipo redondeado, y las anteriores /i, e/, no redondeadas, tengan mayor dificultad de identificación entre ellas cuando los datos visuales y auditivos estén cruzados dentro de cada grupo.

(H3): Las mujeres serán más sensibles a la información procedente del canal visual que los hombres, como indican los resultados señalados por algunos estudios (Argyle e Ingham, 1972; Aloufy *et al*, 1996; Bayliss *et al*, 2005; Traunmüller y Öhrström, 2007).

La confección de los estímulos de este experimento se ha hecho exclusivamente sobre vocales, de modo que se pidió al informante que pronunciara las cinco vocales por separado, es decir, sin contexto. Dicho esto, es preciso tener en cuenta no obstante que desde una perspectiva acústico-perceptiva algunos autores han señalado que las vocales pueden ser percibidas mejor acompañadas de consonantes que de forma aislada, por las importantes pistas que aportan los márgenes consonánticos en los procesos de transición coarticulatoria. Así, en la bibliografía se encuentran estudios que muestran cómo los segmentos vocálicos se identifican mejor durante las transiciones CV o VC, o en contextos silábicos CVC (normalmente oclusivas y siempre las mismas en posición de ataque y coda) que a partir de su núcleo, aunque sus valores acústicos sean más estables (Strange *et al*, 1976; 1983). En esta línea, Jenkins *et al.* (1983), en un estudio sobre percepción de estímulos CVC, mostraron cómo era posible «borrar» la mayor parte de la porción vocálica y seguir manteniendo la identificación del segmento, aunque por otra parte la identificación de la vocal alcanzaba también resultados igualmente buenos si solo se aportaba la información espectral redundante del centro vocálico. Estos mismos resultados se alcanzaron incluso cuando los oyentes eran capaces de identificar vocales a partir de únicamente una breve porción de esa transición o también del propio centro vocálico (Gail *et al,* 2010).

Considerando que el español es una lengua con un sistema vocálico relativamente reducido y simétrico comparado con otros idiomas, y que sus vocales ocupan espacios acústicos amplios, se espera que la identificación aislada de las frecuencias canónicas de sus componentes, en condiciones unimodales, no genere aquí ninguna repercusión perceptiva. En cualquier caso, para observar el comportamiento de esta variable y analizar, como se ha visto, el grado de transmisión informativa de cada señal por separado, se ha efectuado paralelamente un análisis de percepción unimodal de cada una de las señales.

3. METODOLOGÍA

3.1. Participantes

Un total de 28 personas participó como voluntaria en el presente estudio, contabilizando 12 hombres y 16 mujeres. La edad de los participantes oscilaba entre los 33 y los 53 años. Todos eran hablantes nativos de español, concretamente de la variedad dialectal castellana. Ninguno de los participantes tenía problemas de audición, y su visión era también normal, en algunos casos corregida mediante el uso de lentillas o gafas. Los participantes procedían de diversos campos profesionales: administración, enseñanza, sanidad y sector servicios.

La división de los participantes en dos grupos diferenciados por sexo viene motivada por estudios que señalan niveles más altos de atención visual en mujeres que en hombres, pues se ha observado que las mujeres miran al rostro del interlocutor de forma más frecuente y durante periodos más largos de tiempo[11] (Argyle e Ingham, 1972; Bayliss *et al,* 2005) lo que a la postre proporciona mejores niveles de rendimiento en lectura labiofacial (Johnson *et al,* 1988). Es razonable asumir por tanto que el hecho de que las mujeres sean más atentas a la señal visual del habla puede afectar a la sensibilidad de ese canal en el proceso de percepción bimodal, y que esto ofrezca resultados distintos con respecto a los hombres (Traunmüller y Öhrström, 2007, para hablantes de sueco). Este hecho parece confirmarse también con hablantes de inglés (y en menor medida de hebreo) en Aloufy *et al.* (1996), donde las mujeres reflejaban una mayor inclinación por la señal visual que los hombres cuando las claves audiovisuales eran discordantes. Por otra parte, algunos estudios que investigan el procesamiento del lenguaje a partir del análisis neurofisiológico que ofrecen las tomografías computarizadas (TC) o las imágenes por resonancia magnética (RM), muestran evidencias indirectas que justifican posibles diferencias entre hombres y mujeres (Irwin *et al,* 2006), aunque al mismo tiempo estos resultados pueden variar en función del tipo de tarea y su dificultad. Con el objeto de estudiar los resultados en ambos sexos, se intentó reclutar a un número representativo de hombres y mujeres y analizarlos por separado.

La consideración de un rango de edad entre los 33 y 53 años viene motivada a su vez por razones de control de variables. Se sabe que el efecto McGurk aparece en

[11] Este fenómeno se ha documentado también en edades muy tempranas, con bebés de entre 13 y 18 semanas (Leeb y Rejskind, 2004).

todas las edades, aunque se producen diferencias de sensibilidad hacia la señal auditiva o visual en función del rango de edad estudiado. En el estudio de McGurk y MacDonald (1976), se observó que cuando dos señales se presentan de forma incongruente y predomina una de las dos, esta suele ser visual para los adultos (produciendo por tanto un efecto McGurk más fuerte) y auditiva para niños y jóvenes. Esos resultados fueron confirmados también por otros autores (Dodd, 1979; Kuhl y Meltzoff, 1982; Massaro *et al*, 1986), y se relacionan con mecanismos de experiencia y propiocepción en la articulación de sonidos que se desarrollan con la edad, lo que aumenta con la práctica la sensibilidad hacia la información que ofrece el canal visual. Por otra parte, en el extremo opuesto adultos muy mayores pueden reflejar resultados desvirtuados por posibles problemas auditivos y visuales. Para evitar esto, en el presente experimento se descartan niños, pero también adultos cuya edad aumente o disminuya demasiado el rango de edad (20 años) analizado.

Varias consideraciones finales sobre el control de variables que pueden surgir tras la elección de participantes tienen que ver con la lengua y la cultura. Se sabe que la señal visual tiene menos influencia en la percepción del habla en algunas culturas que en otras, como ocurre entre la japonesa y la estadounidense (Sekiyama y Tohkura, 1993). Estas diferencias tienen una fuerte base cultural, ya que por ejemplo en Japón es común que los niños sean transportados en la espalda de la madre, con lo que no adquieren un fuerte lazo visual, y en la propia cultura existen normas de no mirar a la cara del interlocutor durante la interacción oral (Morain, 2001). Esto implica que en los experimentos realizados con hablantes de lenguas que no buscan el contacto visual los resultados del efecto McGurk sean muy pobres, aunque se comprobó que aumentaban significativamente cuando se añade ruido (Sekiyama y Tohkura, 1991), como señalaba de forma recurrente la bibliografía. A ello se añade que hablantes de lenguas con inventarios fonológicos más simples (por ejemplo el japonés frente al inglés) suelen usar un procesamiento perceptivo del habla más dependiente del canal auditivo (Sekiyama *et al*, 2003), que es lo que también se espera en este experimento.

Por otra parte, las variaciones en el modo en que se produce la integración audiovisual también aparecen cuando el hablante no es nativo (Reisberg *et al*, 1987; Chen y Hazan, 2007), ya que se enfatiza más la información que llega por el canal visual en los procesos de percepción bimodal de hablantes que usan el idioma como lengua extranjera que en los procesos perceptivos de L1. Para evitar esto, en el presente experimento tanto el hablante como los participantes del estudio perceptivo comparten una misma lengua y una misma variedad dialectal.

Por último, ninguno de los participantes tiene relación alguna con el hablante. Análisis comparativos realizados entre grupos que en un caso tienen familiaridad con el hablante y en otro no, muestran que cuando la señal audiovisual es incongruente, los participantes familiarizados con el rostro del hablante son menos susceptibles a la señal visual (Walker *et al*, 1995). Para controlar la independencia de estas variables, se comprobó que ninguno de los participantes conocía al locutor.

3.2. Estímulos

Los estímulos corresponden a una grabación estéreo de vídeo y audio de las cinco vocales del español. La selección del material se hizo buscando que los estímulos fueran lo suficientemente marcados como para poder distinguirse entre ellos. Se hicieron por tanto varias pruebas, y se seleccionaron los más representativos desde un punto de vista acústico (timbre estable) y visual (gestos articulatorios bien definidos para cada segmento). Como señalan Summerfield y McGrath (1984:54), frente a la articulación consonántica por lo general breve, acompañada por un mayor número de variaciones en el espectro acústico, la articulación vocálica conlleva movimientos más lentos que se extienden durante más tiempo. En este mismo sentido Gil Fernández (2007:426) añade que las vocales se caracterizan, entre otros aspectos, por la mayor estabilidad de las posiciones articulatorias. Acústicamente también son diferentes: no presentan ruidos aperiódicos, poseen una frecuencia fundamental más alta, son siempre sonoras, duran por lo general más tiempo, y son más audibles o perceptibles que las consonantes, lo que se refleja en un espectro acústico más estable[12]. La carta de formantes y los valores acústicos de los segmentos analizados pueden consultarse en la tabla 1 y en la figura 1.

	i	e	a	o	u
F1	302	504	849	548	324
F2	1918	1869	1259	821	687

Tabla 1. *Valores de F1 y F2 de las vocales del informante.*

[12] Es evidente por otra parte que todo sonido vocálico se ve afectado por factores como la velocidad de articulación, el acento o la coarticulación, y que, desde el punto de vista de la señal visual, además de la velocidad de articulación pueden ocurrir fenómenos de compensación articulatoria que disminuyen la potencia informativa de este canal.

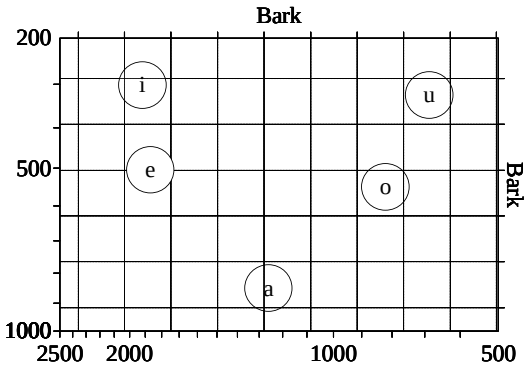


Figura 1. *Carta de formantes de las vocales del informante.*

3.3. Grabación y tratamiento de las vocales

La grabación se realizó en una habitación sin ruido con una cámara Panasonic HC-V500. Tanto la cámara como el rostro del hablante se encontraban a la misma altura, a un metro y medio de distancia. La imagen de su rostro era completa.

El hablante era un hombre de 33 años. Hablaba español en la variedad dialectal castellana, no llevaba gafas y no tenía pelo facial. Se sentó frente a un fondo blanco y se le pidió que pronunciara las cinco vocales del español /a, e, i, o, u/ en diferentes órdenes, intentando que cada una de ellas respetara su gesto articulatorio, con el fin de facilitar la identificación visual de las mismas (ver figura 2). En este sentido se dieron también instrucciones para que las vocales tuvieran una misma duración aproximada entre ellas, que resultó ser de en torno a los 500 ms para cada segmento. Cada serie de vocales fue repetida varias veces para tener un amplio banco de muestras. Finalmente se eligieron los cinco mejores prototipos que reunían las características buscadas en estos estímulos.

Una vez realizada la selección de las cinco vocales, estas fueron tratadas y distribuidas en tres bloques: a) audiovisual; b) solo vídeo; c) solo audio. Por supuesto los cinco segmentos vocálicos analizados eran siempre los mismos en cualquier bloque. El bloque audiovisual constó de 25 combinaciones cruzadas entre todas las vocales, dispuestas en orden aleatorio. Esto permitía combinar cada vocal en cualquier modalidad con todas las vocales. La segunda parte (solo vídeo)

consta de 5 estímulos dispuestos también en orden aleatorio y repetidos dos veces, contabilizándose por tanto un total de 10 estímulos (5 x 2). La repetición permite replicar el proceso de identificación de estos estímulos a partir de la lectura labiofacial, y evaluar si la selección de los mismos se mantiene. Finalmente la tercera parte (solo audio) consta de cinco estímulos (uno para cada vocal). Esto permite controlar que la percepción unimodal de este tipo es fuerte, y no se ve alterada por la ausencia de un mayor marco contextual. A su vez, con el fin de observar mejor el grado de seguridad en la selección de la respuesta, todos los estímulos deben ser evaluados en función del grado de certeza, distribuido en tres niveles (sí / quizá / no). Se evalúan por tanto 40 estímulos vocálicos en 28 sujetos, contabilizándose 1120 respuestas y 3 niveles de certeza por cada una de ellas.

Todos los estímulos fueron separados, tratados y combinados usando el programa Pinnacle Studio V15. Cada estímulo audiovisual y visual duraba alrededor de 3 segundos, y siempre comenzaba y terminaba con la imagen del hablante en posición de reposo y con la boca cerrada. No hubo problemas de sincronización de audio y vídeo en los estímulos audiovisuales discordantes de la primera parte. En el último bloque del experimento (solo estímulos auditivos) la pantalla estaba en negro.

Figura 2. *Distintas secuencias del test utilizado.*

3.4. Procedimiento experimental

Todos los sujetos participaron de uno en uno. Los estímulos se presentaron a partir de la pantalla y altavoz integrados de un ordenador Acer TravelMate 292 LMI, de 15", colocado a unos 60 cm. Se informó a los participantes de que iban a ver un vídeo en el que aparecía un hablante pronunciando las cinco vocales del español. Para realizar el test perceptivo se les pidió que miraran con atención los gestos articulatorios del hablante, y se subrayó que este aspecto era muy importante (Paré *et al,* 2003). Con el objeto de que pudieran escribir sus respuestas, se proporcionó en papel una copia del test. En este punto se insistió también sobre la necesidad de escribir una única respuesta por estímulo. Es preciso destacar que la propia pregunta del test evitaba predisponer al sujeto a favor de una modalidad o de otra (Colin *et al,* 2005:22), por lo que se evitaron cuestiones del tipo («¿qué vocal oyes?» o «¿qué vocal ves?»). La pregunta que aquí se formuló fue («¿qué vocal entiendes?») y se comunicó tanto de forma oral como a través de las instrucciones del propio vídeo antes de cada bloque.

El vídeo duró un total de 8,38 minutos. Con el fin de que los participantes tuvieran tiempo de anotar sus respuestas, los estímulos estaban separados por un margen de 8 segundos, al final de los cuales aparecía un pitido que anunciaba el comienzo del siguiente estímulo. Todos los sujetos fueron supervisados en cada sesión para asegurar que miraban a la pantalla todo el tiempo.

4. RESULTADOS

Con el fin de facilitar el seguimiento, se muestran a continuación los resultados estructurados por bloques, comparando los grupos de jueces cuando así procede. El análisis estadístico se ha realizado mediante la aplicación SPSS Statistics 20, fijando un nivel de significación del 5% ($p<0,05$). Como herramientas estadísticas se han utilizado las tablas de contingencia con test Chi-cuadrado de independencia entre dos variables cualitativas, y la correlación de Spearman. Los gráficos se han obtenido a través del programa GraphPad Prism 6.

4.1. Bloque primero (serie audiovisual)

El objetivo principal de este bloque es comprobar si para cada una de las 5 vocales del español en todos sus cruces audiovisuales prima el estímulo auditivo frente al

visual, para lo cual se utilizarán contrastes de hipótesis sobre la diferencia de proporciones. Después de extraer una conclusión global se contrastará si la percepción auditiva es también la prioritaria para hombres y para mujeres, tratados esta vez de forma independiente, y se intentará determinar si se aprecian diferencias significativas entre las percepciones auditivas y visuales entre ambos sexos. Por último se expondrán de forma más detallada los resultados obtenidos en cada uno de los 25 estímulos cruzados dentro del grupo de hombres y de mujeres teniendo en cuenta la respuesta ante el estímulo auditivo, se describirán estos cruces a partir de las respuestas ante los estímulos visuales, y se estudiará el comportamiento de las fusiones, es decir, cuáles son las características de las vocales elegidas cuando estas no se corresponden ni con el canal auditivo ni con el visual.

Con el fin de calcular la prevalencia de un canal frente al otro, se tomaron en cuenta únicamente las 20 combinaciones audiovisuales de vocales discordantes (por tanto, no a-a; e-e; i-i; o-o; u-u) y se calcularon los porcentajes de acierto en las dos señales. Para ello se consideró: a) como acierto auditivo el porcentaje de veces que los jueces acertaron la vocal que escuchaban sobre el total (es decir, aquellas respuestas que coincidían con la señal auditiva); b) como acierto visual el porcentaje de veces que los jueces acertaron la vocal que habían visto pronunciar (por tanto, cuyas respuestas coincidían con la señal visual); y c) como fusión el porcentaje de jueces que no seleccionaron ni la vocal que escuchaban ni la que veían en el momento del experimento.

Como se observa en la figura 3, el porcentaje de acierto auditivo es en general bastante alto (81%), y considerablemente mayor que el acierto visual, un 12%. Estas diferencias son estadísticamente significativas, con $p<,05$ ($p=,001$). Existe también un porcentaje residual de jueces (7%) que no han seleccionado la vocal que les llegaba por uno u otro canal. Estos resultados permiten avalar por tanto la existencia de un efecto McGurk para esta clase de estímulos y condiciones, ya que en un 19% de los casos la incongruencia de los estímulos audiovisuales compromete la percepción de la señal auditiva.

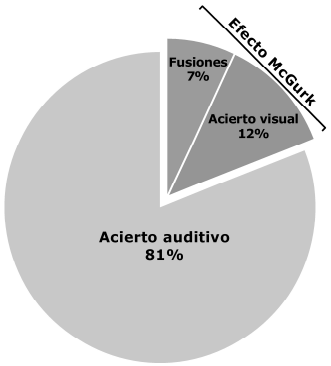


Figura 3. *Distribución del porcentaje de aciertos y fusiones.*

Si prestamos ahora atención a los porcentajes de acierto auditivo para cada una de las vocales (figura 4) parece bastante llamativo cómo las tres primeras /a, e, i/ presentan un porcentaje de acierto muy alto, con valores de 89,3%, 91,1% y 87,5% respectivamente, frente al 75,9% para la vocal /o/ y del 60,7% para la vocal /u/. Como se aprecia en la tabla 2, la prueba Chi-cuadrado evidencia que estas diferencias en la proporción de aciertos de /o/ y /u/ con respecto a /a, e, i/ son estadísticamente significativas (Chi2 = 47,09; 4 gl ; p<0,001).

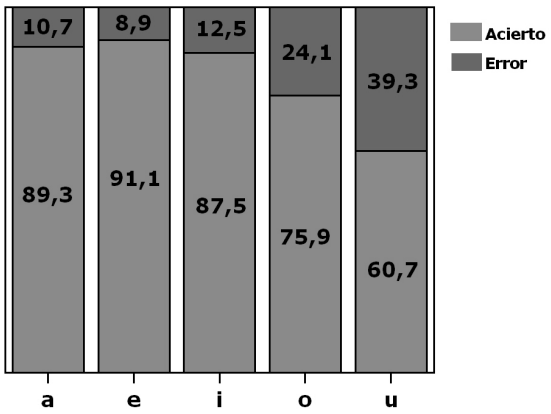


Figura 4. *Porcentajes de acierto y error auditivos.*

La razón de estos resultados podría deberse a dos factores íntimamente relacionados. Por un lado, es preciso tener en cuenta que desde un punto de vista acústico-perceptivo las vocales con frecuencias más altas, como /i, e, a/, presentan respectivamente los grados de «claridad» más elevados, de modo que perceptivamente tienden a oscurecerse a medida que en el espectro van disminuyendo sus frecuencias. Esto hace que las vocales anteriores sean más claras que las posteriores, y las no redondeadas más claras que las redondeadas (Gil, 2007). Considerado esto, tanto la vocal /o/ como especialmente la vocal /u/ se caracterizan por tener un timbre grave, ya que en ellas predominan las frecuencias bajas, especialmente si, como es el caso, son redondeadas. Estas características acústico-perceptivas podría ponerlas en especial desventaja en este tipo de situaciones audiovisuales incongruentes.

Unido a ello, es necesario considerar también las particulares características acústicas de estos dos segmentos, que en este experimento presentan unos valores de F2 ligeramente bajos, seguramente motivados por producirse con un gesto articulatorio más marcado que, con el fin de ayudar a visualizar mejor el redondeamiento labial de estos sonidos, origina valores de F3 y F2 más bajos (Ladefoged y Johnson, 2006). Este aspecto parece reflejarse con especial intensidad en la /u/, que es precisamente la vocal que presenta articulatoriamente un redondeamiento más marcado. Aunque los valores entran dentro de los campos de dispersión vocálica marcados para el español por algunos estudios (Martínez Celdrán y Fernández Planas, 2007), y las vocales son reconocidas perceptivamente al 100% por todos estos mismos hablantes en condiciones unimodales, como después se verá, es posible que en los cruces audiovisuales incongruentes las pequeñas variaciones acústicas alejadas de sus «centros», especialmente en el caso de la /u/, que en este experimento presenta valores más extremos, sean más sensibles a repercusiones perceptivas en estas condiciones.

	Vocal n (%)				
Resultado	a	e	i	o	u
Acierto	100a (89,3)	102a (91,1)	98a (87,5)	85b (75,9)	68c (60,7)
Error	12a (10,7)	10a (8,9)	14a (12,5)	27b (24,1)	44c (39,3)

Chi-cuadrado: valor=47,09; gl=4; p<0,001

Tabla 2. *Prueba Chi-cuadrado resultado auditivo por vocal (condiciones audiovisuales).*

Si tenemos ahora en cuenta el acierto visual de las vocales, tal y como puede apreciarse en la figura 5, el comportamiento se presenta prácticamente a la inversa que el auditivo. Aumenta así el porcentaje de aciertos visuales en las vocales /a, o, u/ frente al de las demás, por lo que parece que el grado de abertura máximo y la presencia de redondeamiento labial ofrecen una destacada información visual que las hace sobresalir cuando se elige la información que llega por este canal.

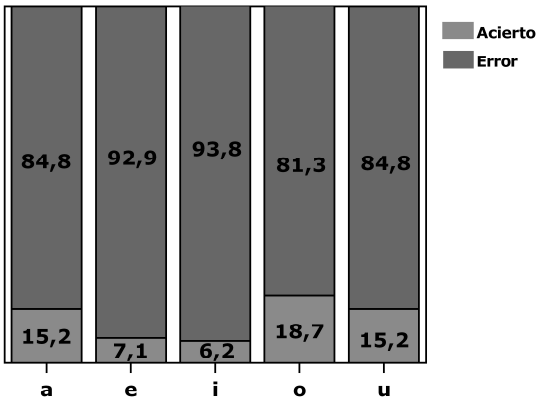


Figura 5. *Porcentajes de acierto y error visuales.*

Como se aprecia en la tabla 3, al comparar los porcentajes de acierto y error de la prueba se han encontrado diferencias estadísticamente significativas entre unas vocales y otras (Chi2= 12,41; 4 gl; p=0,015). Así, la vocal /i/, que es la vocal con menores porcentajes de identificación visual, presenta diferencias significativas con todas las vocales excepto /e/, mientras que la /o/, que es la vocal que presenta mayores porcentajes de identificación, presenta diferencias significativas con respecto a /e, i/, pero no con respecto a /a, u/. Se aprecia pues una clara diferencia de aciertos visuales entre el grupo de vocales que presenta mayores claves visuales (/a, o, u/), y las vocales que no (/e, i/), especialmente esta última. Por otra parte, como puede apreciarse en la figura 6, el porcentaje de aciertos visuales sigue siendo considerablemente menor que el porcentaje de aciertos auditivos. En todas las vocales estas diferencias son estadísticamente significativas (χ_1^2=10,63; p=0,001).

	Vocal n (%)				
Resultado	a	e	i	o	u
Acierto	17a, b (15,2)	8b, c (7,1)	7c (6,2)	21a (18,7)	17a, b (15,2)
Error	95a, b (84,8)	104b, c (92,9)	105c (93,8)	91a (81,3)	95a, b (84,8)

Chi-cuadrado: valor=12,41; gl=4; p=0,015

Tabla 3. *Prueba Chi-cuadrado resultado visual por vocal (condiciones audiovisuales).*

En cuanto a las posibles diferencias entre hombres y mujeres, tal y como puede verse en la figura 6 la percepción auditiva en ambos grupos es superior a la percepción visual, con diferencias en ambos estadísticamente significativas (χ_1^2=10,78; p=0,001). Considerado esto, se analizó también si pueden observarse diferencias entre hombres y mujeres en relación con el porcentaje de acierto auditivo y acierto visual. Como puede apreciarse de nuevo en la figura 6, cuando se presentan estímulos en condiciones audiovisuales incongruentes los hombres presentan un mayor nivel de acierto en el estímulo auditivo que las mujeres (86,25% frente al 77,18%), que es estadísticamente significativo (χ_1^2=6,34; p=0,01). Por su parte, las mujeres tienen más acierto que los hombres en la identificación de estímulos que llegan del canal visual (14,4% frente a 9,6%), siendo esta diferencia también estadísticamente significativa (χ_1^2=4,13; p=0,042). Estos resultados parecen apuntar a una mayor sensibilidad de las mujeres hacia la señal visual, tal como muestran algunos estudios (Aloufy *et al*, 1996; Bayliss *et al*, 2005; Traunmüller y Öhrström, 2007), aunque es importante reseñar que el pequeño tamaño de la muestra no permite concluir aquí que estas diferencias sean extrapolables a la población general.

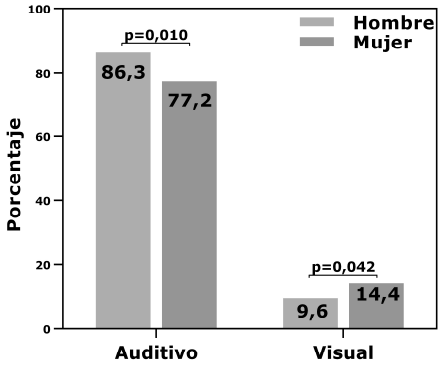


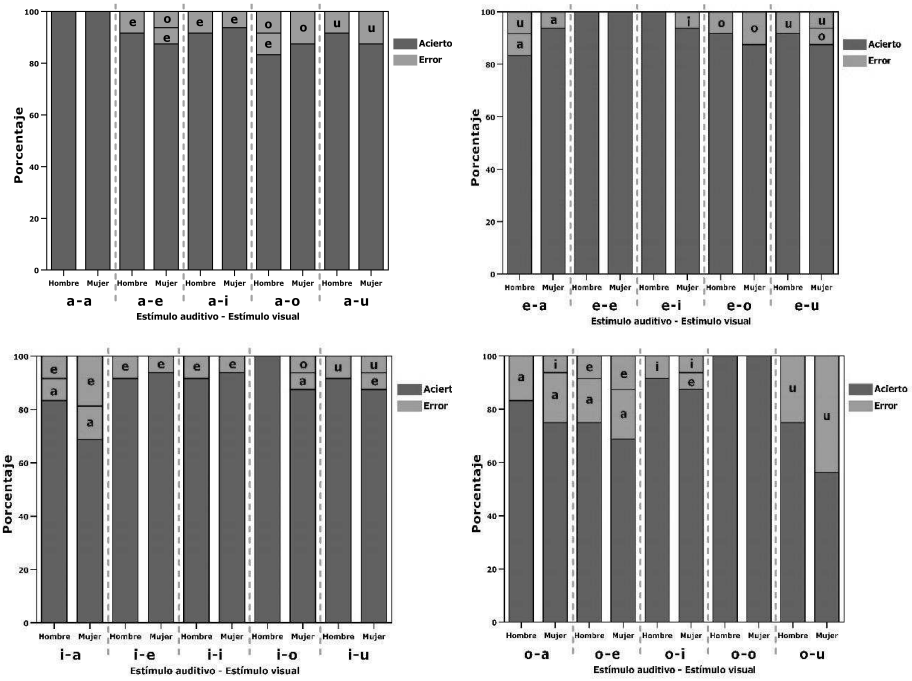
Figura 6. *Porcentajes de acierto auditivo y visual por sexo.*

A continuación, en la tabla 4 se resumen en calidad porcentual los resultados obtenidos en cada uno de los 25 estímulos cruzados en el grupo de los hombres y de las mujeres, teniendo como referencia el acierto auditivo, es decir, las respuestas de los hablantes que coinciden con este estímulo. La representación pormenorizada de estas respuestas por cada una de estas combinaciones audiovisuales puede encontrarse a su vez en la figura 7, donde se representan las características de los aciertos y errores tomando igualmente como referencia el acierto auditivo.

Hombres		ESTÍMULO AUDITIVO				
		i	e	a	o	u
ESTÍMULO VISUAL	i	91,67%	100%	91,67%	91,67%	75%
	e	91,67%	100%	91,67%	75%	83,33%
	a	83,33%	83,33%	100%	83,33%	83,33%
	o	100%	91,67%	83,33%	100%	66,67%
	u	91,67%	91,67%	91,67%	75%	100%

Mujeres		ESTÍMULO AUDITIVO				
		i	e	a	o	u
ESTÍMULO VISUAL	i	93,75%	93,75%	93,75%	87,50%	50%
	e	93,75%	100%	87,50%	68,75%	56,25%
	a	68,75%	93,75%	100%	75%	50%
	o	87,50%	87,50%	87,50%	100%	37,50%
	u	87,50%	87,50%	87,50%	56,25%	93,75%

Tabla 4. *Distribución porcentual de las respuestas tomando como referencia el acierto auditivo (n hombres = 12; n mujeres= 16).*



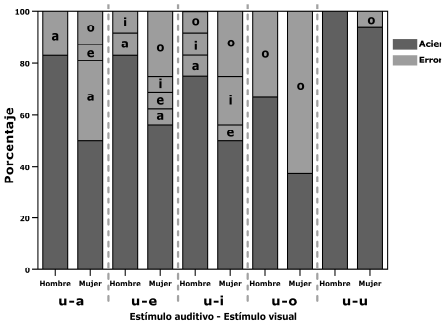


Figura 7. *Representación completa de las respuestas a partir de los estímulos auditivos.*

A la vista de estos datos cabe señalar que:

1. La respuesta ante la concordancia de estímulo visual y auditivo (i-i, e-e, a-a, o-o, u-u), cuya identificación se habría esperado en el 100% de los casos, se produce en el caso de los hombres en todos los pares a excepción del i-i; en el caso de las mujeres además del par i-i tampoco se produce en el par u-u. El hecho de que no se haya obtenido un 100% de aciertos en los cinco pares vocálicos congruentes, como se esperaba, puede deberse a la sensibilidad de los jueces a las condiciones audiovisuales discordantes de este bloque. Casi todos los jueces manifestaron que se daban cuenta de la incongruencia entre ambas señales, por lo que es posible que desarrollaran una predisposición a percibir una manipulación en todos los estímulos presentados, incluso cuando no era así. En cualquier caso, en los pares congruentes donde no hay unanimidad de acierto, los porcentajes en los dos grupos están por encima del 90%.

2. El grupo de las mujeres parece estar más influenciado por la señal visual, pues en la mayoría de los casos este grupo ofrece menores porcentajes de acierto auditivo que el grupo de los hombres, concretamente en 16 de las 20 combinaciones incongruentes. Estos porcentajes más bajos, con resultados del 75% (e incluso menores), aparecen fundamentalmente en todas las combinaciones en las que intervienen las vocales /o/ y /u/ auditivas, siendo especialmente reseñable, con un 37,50% de acierto auditivo, la respuesta al estímulo *auditivo u/visual o,* lo que convierte al grupo de mujeres en el único que en al menos un caso no opta mayoritariamente por la señal auditiva. En el muestreo realizado en los hombres

los pares más conflictivos resultaron ser el estímulo *auditivo u/visual i*, el *estímulo auditivo o/visual e*, el *estímulo auditivo o/visual u* (todos con un 75%), y especialmente alejado de la respuesta esperada, el estímulo *auditivo u/visual o* (66,67%). Resulta llamativo por otra parte que en el grupo de los hombres existen dos pares incongruentes donde la respuesta es 100%, como son el estímulo *auditivo e/visual i*, y el estímulo *auditivo i/visual o*, donde la totalidad de los hombres elige el estímulo auditivo, frente al 93,75% y el 87,5% respectivamente en el caso de las mujeres. De este modo solamente el grupo de los hombres alcanza alguna vez resultados del 100% de acierto auditivo cuando los estímulos audiovisuales son incongruentes.

3. El grupo de las mujeres ofrece también mayor variabilidad en la respuesta. Esto sucede así en 10 de 25 combinaciones, frente a solamente 2 combinaciones de 25 en las que los hombres eligen un mayor número de respuestas que las mujeres. En las 13 combinaciones restantes la variabilidad es la misma. En este sentido, resultan especialmente llamativos los estímulos *auditivo u/visual a*, donde las mujeres eligen hasta tres vocales diferentes a la auditiva, frente a solamente una vocal por parte de los hombres, y *auditivo u/visual e*, donde las mujeres eligen cuatro vocales distintas al estímulo auditivo, frente a solamente dos en el caso de los hombres.

4. Cuando las vocales posteriores /o, u/ se presentan en audio, los jueces muestran más variabilidad, especialmente el grupo femenino. Como se aprecia en la tabla 5, el porcentaje de aciertos en la vocal /u/ en los hombres es del 77,1%, frente a un 48,4% en las mujeres. La prueba Chi-cuadrado evidencia que ambas variables son dependientes, y por tanto se puede afirmar que el porcentaje de aciertos en hombres es mayor que en mujeres (χ^2_1=9,44; p=0,002). Para el resto de vocales, el porcentaje de aciertos entre hombres y mujeres no presenta diferencias estadísticamente significativas: /a/ (χ^2_1=0,01; p=0,930); /e/ (χ^2_1=0,04; p=0,848); /i/ (χ^2_1=1,33; p=0,248); /o/ (χ^2_1=1,32; p=0,251). Como se vio, la razón de estos resultados podría deberse a las especiales características acústicas de la /u/, que, al igual que sucedió con la /o/ en menor medida, en este experimento presentan unos valores de F2 ligeramente bajos y alejados de sus centros, lo que a la postre refuerza las características de un timbre que en estas vocales ya es de por sí más oscuro. Aunque estas variaciones en la frecuencia canónica de sus formantes no tienen ninguna repercusión perceptiva en condiciones unimodales, y están dentro de los límites de sus campos de dispersión señalados por la bibliografía para cada una de estas vocales, es posible que la mayor sensibilidad del grupo femenino hacia la señal visual redujera el acierto auditivo en las combinaciones en las que intervienen estos estímulos auditivos.

		Sexo n(%)	
Vocal	Resultado	Hombre	Mujer
a	Acierto	43 (89,6)	57 (89,1)
	Error	5 (10,4)	7 (10,9)
e	Acierto	44 (91,7)	58 (90,6)
	Error	4 (8,3)	6 (9,4)
i	Acierto	44 (91,7)	54 (84,4)
	Error	4 (8,3)	10 (15,6)
o	Acierto	39 (81,3)	46 (71,9)
	Error	9 (18,8)	18 (28,1)
u	Acierto	37 (77,1)	31 (48,4)
	Error	11 (22,9)	33 (51,6)

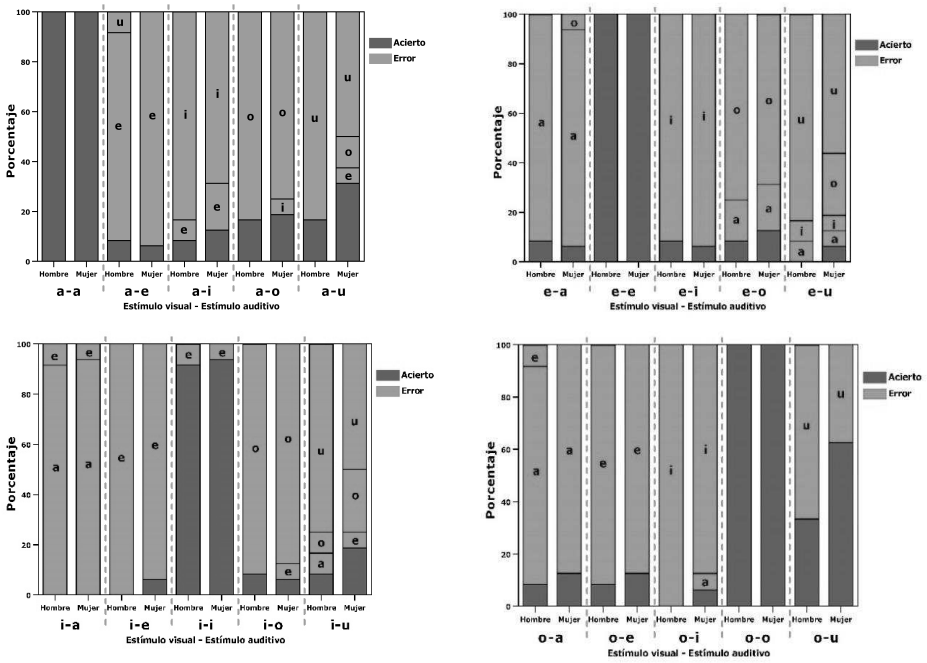
Tabla 5. *Prueba Chi-cuadrado resultado auditivo por vocal y sexo en condiciones audiovisuales incongruentes.*

La tabla 6 presenta a continuación una visión de los datos teniendo en este caso como referencia el acierto visual. La representación pormenorizada de estas respuestas por cada una de estas combinaciones audiovisuales puede encontrarse a su vez en la figura 8, donde se representan las características de los aciertos y errores teniendo igualmente como referencia el acierto visual.

Hombres		ESTÍMULO AUDITIVO				
		i	e	a	o	u
ESTÍMULO VISUAL	i	-	8,33%	8,33%	0%	8,33%
	e	0%	-	8,33%	8,33%	8,33%
	a	0%	8,33%	-	8,33%	8,33%
	o	8,33%	8,33%	16,6%	-	25%
	u	8,33%	0%	16,6%	33,3%	-

Mujeres		ESTÍMULO AUDITIVO				
		i	e	a	o	u
ESTÍMULO VISUAL	i	-	6,25%	12,5%	6,25%	6,25%
	e	6,25%	-	6,25%	12,5%	6,25%
	a	0%	6,25%	-	12,5%	6,25%
	o	6,25%	12,5%	18,75%	-	43,75%
	u	18,75%	6,25%	31,25%	62,5%	-

Tabla 6. *Distribución porcentual de las respuestas tomando como referencia el acierto visual (n hombres = 12; n mujeres= 16).*



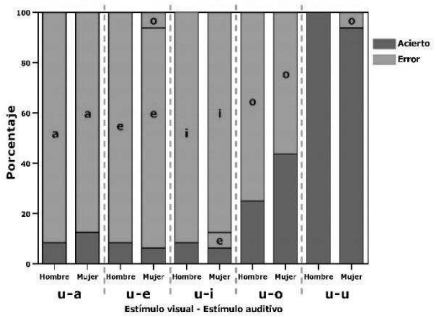


Figura 8. *Representación completa de las respuestas a partir de los estímulos visuales.*

A la vista de estos resultados cabe señalar que:

1. Como se vio en la tabla 6, la /o/, la /a/ y la /u/ ofrecen por este orden los mejores porcentajes de acierto visual tanto en hombres como en mujeres, especialmente en combinación con /o/ y /u/ auditivas. Esto permite afirmar que el grado de abertura máximo y la presencia de redondeamiento labial son las dimensiones más influyentes de la señal visual de las vocales del español con esta serie de estímulos y condiciones. Por otra parte, en los dos grupos las vocales posteriores /o, u/, presentadas de forma visual, ofrecen altos porcentajes de confusión cuando se combinan entre ellas, especialmente para el grupo de mujeres, que también es más susceptible al redondeamiento que ofrece la señal visual cuando /o, u/ se combinan con otras vocales. No ocurre lo mismo con las vocales anteriores /i, e/, las cuales parecen tener más en cuenta la señal auditiva cuando los estímulos están cruzados.

2. En todas las combinaciones aparece al menos un hablante que ha elegido el estímulo visual, excepto el par *visual i / auditivo a*, donde ni hombres ni mujeres se han decantado por la señal visual. Si se observa ahora el comportamiento de cada grupo, se puede apreciar que el grupo de mujeres elige la señal visual en un mayor número de pares audiovisuales que el grupo de los hombres, el cual presenta tres combinaciones sin que haya una elección por el estímulo visual, frente a solo un caso por parte del grupo de mujeres. Por otra parte, el grupo de mujeres presenta también el único caso, con un 62,5%, en el que hay una mayor preferencia por el estímulo visual que por el auditivo (*visual o / auditivo u*).

3. El grupo de las mujeres ofrece mayores porcentajes de acierto visual que el grupo de los hombres, concretamente en 13 de las 20 combinaciones incongruentes. Estos porcentajes más altos ocurren sistemáticamente en todas las combinaciones en las que la vocal visual es una /o/, siendo la combinación cruzada estímulo *visual o/auditivo u* el caso en el que las diferencias porcentuales son mayores. Por su parte, el grupo de los hombres ofrece porcentajes más altos de acierto visual en 6 de 20 combinaciones audiovisuales incongruentes (la combinación restante, como se señalaba en el punto anterior, se corresponde con el par *visual i / auditivo a*, en la que ni hombres ni mujeres eligen el estímulo visual). Considerado esto, si se analiza el porcentaje de aciertos y errores para cada vocal entre hombres y mujeres (ver tabla 7), no se encuentran diferencias estadísticamente significativas en ninguna vocal: /a/ (χ^2_1=0,47; p=0,494); /e/ (χ^2_1=0,10; p=0,751); /i/ (χ^2_1=0,62; p=0,430); /o/ (χ^2_1=2,15; p=0,142); /u/ (χ^2_1=0,47; p=0,494).

		Sexo n(%)	
Vocal	Resultado	Hombre	Mujer
a	Acierto	6 (12,5)	11 (17,2)
	Error	42 (87,5)	53 (82,8)
e	Acierto	3 (6,3)	5 (7,8)
	Error	45 (93,8)	59 (92,2)
i	Acierto	2 (4,2)	5 (7,8)
	Error	46 (95,8)	59 (92,2)
o	Acierto	6 (12,5)	15 (23,4)
	Error	42 (87,5)	49 (76,6)
u	Acierto	6 (12,5)	11 (17,2)
	Error	42 (87,5)	53 (82,8)

Tabla 7. *Prueba Chi-cuadrado resultado visual por vocal y sexo en condiciones audiovisuales incongruentes.*

4. Como se aprecia en la figura 8, las mujeres ofrecen de forma consistente una mayor variabilidad de respuestas que los hombres frente a los mismos estímulos. Así:

1. En 10 de los 25 estímulos las mujeres registran más respuestas posibles que los hombres. Los casos más notorios son: *estímulo visual e/auditivo u*, donde las mujeres ofrecen como respuesta /a, e, i, o, u/ frente a /u, i, a/ en los hombres; estímulo *visual a/auditivo u*, donde las mujeres ofrecen respuestas /u, a, e, o/ frente a /u, a/ en los hombres; y estímulo *visual o/auditivo i*, donde las mujeres responden /i, o, a/, frente a únicamente /i/ los hombres.

2. Hay 13 ocasiones de 25 donde hombres y mujeres ofrecen igual variabilidad en su respuesta.

3. Hay solamente dos ejemplos en los que los hombres registran una mayor variabilidad, concretamente el estímulo *visual o/auditivo a*, donde ellos responden /a, o, e/, frente a la respuesta de las mujeres /a, o/; y estímulo *visual a/auditivo e*, donde ellos responden /a, e, u/, frente a la respuesta de las mujeres /a, e/.

Finalmente, el tercer análisis de los resultados obtenidos pretende investigar las características del 7% de las respuestas emitidas por los jueces que, según se vio en la presentación de los resultados globales (figura 3), no se correspondían ni con la vocal auditiva ni con la vocal que veían pronunciar. Algunos estudios (Summerfield y McGrath, 1984) muestran que este tipo de fusiones con vocales audiovisuales incongruentes se caracteriza por su tendencia a «desplazar» la cualidad vocálica del segmento auditivo hacia los valores de otra vocal situada en un espacio vocálico contiguo. Con el fin de observar este comportamiento en español, se muestra a continuación un esquema de todos los pares de vocales con sus posibles fusiones resultantes (ver tabla 8).

		Peso porcentual y número de casos			
Audio / vídeo	Fusiones	Porcentaje sobre el total de respuestas (28) por combinación	Nº de casos (hombres)	Nº de casos (mujeres)	Porcentaje sobre el total de fusiones (38) producidas en el experimento
(a / e)	/o/	3,57%	0	1	2,38%
(a / i)	/e/	7,14%	1	1	5,26
(a / o)	/e/	3,57%	0	1	2,38%
(a / u)	/o/	3,57%	0	1	2,38%
(e / a)	/u/	3,57%	1	0	2,38%
(e / i)	No hay fusión	-	-	-	-
(e / o)	No hay fusión	-	-	-	-
(e / u)	/o/	3,57%	0	1	2,38%
(i / a)	/e/	14,28%	1	3	10,52
(i, e)	No hay fusión	-	-	-	-
(i / o)	/a/	3,57%	0	1	2,38%
(i / u)	/e/	3,57%	0	1	2,38%
(o / a)	/i/	3,57%	0	1	2,38%
(o / e)	/a/	17,85%	2	3	13,15%
(o / i)	/e/	3,57%	0	1	2,38%
(o / u)	No hay fusión	-	-	-	-
(u / a)	/e, o/	/e/ = 3,57% /o/= 7,14% *Total: 10,71%*	-	/e/ = 1 /o/ = 2	7,89%
(u / e)	/a, i, o/	/a/ = 7,14% /i/ = 7,14% /o/ = 14,28%	/a/ =1 /i/ = 1	/a/ =1 /i/ =1 /o/= 4	26,31%

		Total: 28,56%			
(u / i)	/a, e, o/	/a/ = 3,57% /e/ = 3,57% /o/ = 17,85% *Total: 25%*	/a/ = 1 /o/ = 1	/a/ = 0 /e/ = 1 /o/ = 4	18,42%
(u / o)	No hay fusión	-	-	-	-
N casos	-	-	9	29	-

Tabla 8[13]. *Casos y porcentajes de fusiones.*

A la vista de estos resultados cabe señalar:

1. Los porcentajes de fusión son especialmente altos en los pares en los que aparece una /u/ auditiva combinada con /e, i/ visuales. De este modo, el par (u, e) presenta un 28,57% de casos de fusión, y el par (u, i) un 25%. Los resultados muestran por tanto que cuando en estas condiciones audiovisuales incongruentes se combinan los estímulos con peores porcentajes de identificación auditiva y los estímulos con peores porcentajes de identificación visual, los porcentajes de fusión resultantes de ambas señales se convierten en los más altos de toda la serie de fusiones. Por otra parte, como puede apreciarse en la tabla 8, de las 20 combinaciones posibles, en 5 no hay fusión (eligen lo visual o lo auditivo). Cuatro de estos casos se producen cuando se combinan, en cualquier orden audiovisual, bien las vocales anteriores (e / i), o bien las vocales posteriores (o / u), en cuyo caso su localización por cualquiera de los dos canales resulta absoluta. Esto mismo ocurre también con el par (e / o), cuyo par «espejo» (o / e) da como único caso de fusión la vocal más próxima a las dos: /a/.

2. Si tenemos en cuenta los casos de fusión por vocal sobre el número total de respuestas[14] (ver tabla 9), se observa que las vocales medias /o/ y /e/ reúnen, por

[13] Para una mayor simplicidad, se presenta la relación auditivo / visual de los pares audiovisuales en este mismo orden, separados por barra, y entre paréntesis, de modo que la primera vocal es siempre la auditiva: ($Vocal^{A}$ / $Vocal^{V}$).

[14] El cálculo se realiza del siguiente modo: 28 sujetos x 20 combinaciones audiovisuales en las que puede haber casos de fusión (por tanto considerando únicamente los estímulos audiovisuales incongruentes) y dividido entre las cinco vocales del español consideradas en este estudio (28 x 20 / 5 = 112 respuestas posibles por vocal).

este orden, el mayor número de casos de fusión (12,5% y 9,8% respectivamente), mientras que las vocales altas /i/ (2,7%) y /u/ (0,9%) son las que reúnen menos. Las diferencias entre ambos grupos de vocales son estadísticamente significativas (Chi2 = 16,83; 4 gl; p= 0,002). La vocal /a/, por su parte, solo presenta diferencias estadísticamente significativas con /u/.

	Vocal n (%)				
Resultado	a	e	i	o	u
Acierto	9a, b, c (8,0%)	11c (9,8%)	3b, d (2,7%)	14a, c (12,5%)	1d (0,9%)
Error	103a, b, c (92,0%)	101c (90,2%)	109b, d (97,3%)	98a, c (87,5%)	111d (99,1%)

Chi-cuadrado: valor=16,83; gl=4; p=0,002

Tabla 9. *Prueba Chi-cuadrado resultados de fusión por vocal sobre el total de respuestas (n=112).*

3. Considerando únicamente los casos de fusión (n= 38), estos se concentran, como se vio en el punto anterior, en las vocales /o/ (36,8%), /e/ (28,9%), y /a/ (23,6%). Por otro lado la /u/ presenta un único caso de fusión (2,6% del total), y hay solamente tres casos para la /i/ (7,8%) (ver figura 9). Como puede apreciarse en la tabla 10, estas diferencias entre las vocales altas y medias se mantienen de un modo estadísticamente significativo.

	Vocal n (%)				
Resultado	a	e	i	o	u
Fusión	9a, b, c (23,7)	11c (28,9)	3b, d (7,9)	14a, c (36,8)	1d (2,6)
No fusión	29a, b, c (76,3)	27c (71,1)	35b, d (92,1)	24a, c (63,2)	37d (97,4)

Chi-cuadrado: valor=19,61; gl=4; p=0,001

Tabla 10. *Prueba Chi-cuadrado resultados de fusión por vocal considerando solo los casos de fusión (n=38).*

Dada la simplicidad del sistema vocálico español, no es sorprendente que los casos de fusión se hayan concentrado en las vocales situadas en la «centralidad» de los ejes del espacio vocálico, sea considerando desde un punto de vista articulatorio la dimensión anteroposterior, o la dimensión de la altura del cuerpo lingual (ver figura 9). Considerado también el reducido número de segmentos que pueden intervenir en la fusión, no es extraño tampoco encontrar pares «espejo» que, independientemente del cruce entre estímulos visuales o auditivos, producen un mismo caso de fusión (aunque en los ejemplos con /u/ auditiva pueden darse también otras fusiones). Los casos a los que se hace referencia son: (a / i; i / a), (a / u; u / a); (e / u; u / e); (i / u; u / i), que representan 8 de las 20 combinaciones posibles, y 8 de las 15 en las que se producen casos de fusión (un 53,3%).

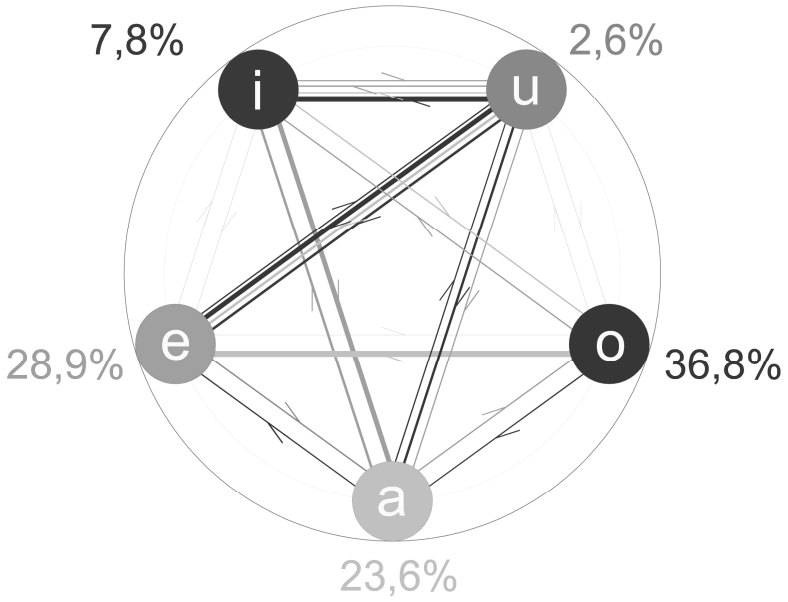


Figura 9. *Porcentajes, distribución y frecuencia de fusiones por vocal (la flecha indica la posición de la vocal auditiva).*

4. Cualquier combinación de una vocal cerrada /i, u/ con /a/ (los extremos del triángulo vocálico) produce como fusión la vocal semicerrada más equidistante de

las dos. Por ejemplo, tal y como se aprecia en la figura 9, los pares (i / a; a / i) dan /e/, y los pares (u / a; a / u) dan /o/ (el par (*u / a)* da también /e/). Otros pares «espejo» que muestran un comportamiento semejante son los pares (e / u) y (u / e), que se fusionan en /o/ (y en el caso de (u / e) también en /i/ y en /a/).

5. Hay cuatro combinaciones donde al menos cuatro personas o más han señalado una fusión hacia una vocal determinada. Son: (i / a) = /e/; (o / e) = /a/; (u / e) = /o/; y (u / i) = /o/, con unos valores de frecuencia de aparición que pueden llegar hasta el 17,85%. Como se aprecia en la figura 9, en todos estos casos las fusiones resultantes se caracterizan por situarse bien en un espacio vocálico contiguo y equidistante al del par audiovisual, o bien en un espacio próximo a la señal auditiva, la cual suele ejercer más influencia que la señal visual a la hora de «desplazar» la cualidad vocálica del segmento resultante, por ser como vimos la señal dominante. Ejemplos de este último caso en el que la fusión se sitúa cerca de la vocal auditiva pueden encontrarse también en (a / o) = /e/; (a / e) = /o/; y (u / i) = /o/.

6. Si se atiende a la relación completa y detallada por vocal de todos los casos de fusión (19 casos distintos en total, ver tabla 11) que surgen de las combinaciones audiovisuales, se puede observar que todas las vocales resultantes de la fusión proceden al menos de un par «inesperado», bien porque esa vocal nueva se sitúa más próxima a la señal visual, o bien porque se sitúa en un extremo opuesto al espacio que ocupa cualquiera de los dos estímulos audiovisuales. Estos ejemplos son: (o / a; u / e) = /i/; (u / a; u / i) = /e/; (u / i; u / e; i / o) = /a/; (e / u) = /o/; y (e / a) = /u/. De estas nueve combinaciones, hay seis pares (u / e) = /i/; (u / a; u / i) = /e/; (u / e; i / o) = /a/; y (e / u) =/o/, que presentan ejemplos de fusiones cuyas vocales se sitúan próximas a la señal visual. Sin embargo ninguna de las fusiones resultantes de los cuatro primeros pares es, como puede verse, la mayoritaria en los pares audiovisuales de los que proceden, más próximos como se dijo anteriormente a la vocal más equidistante de las dos y/o la señal auditiva (en los cuatro casos /o/). Su elección aquí puede deberse a la conjunción de dos factores: por un lado en los cuatro pares el estímulo auditivo se corresponde con la vocal que presenta peores porcentajes de acierto auditivo para este tipo de combinaciones audiovisuales, como es el caso de la /u/. Esto quizá es debido, como se ha señalado, a las propias propiedades acústicas del estímulo presentado, el cual posee unos valores de F2 particularmente bajos por realizarse con un gesto articulatorio especialmente marcado, lo que perceptivamente se manifiesta con un timbre todavía más oscuro. Por otro lado, es preciso considerar también la influencia que puede ejercer la señal visual, así como la mayor sensibilidad hacia esa señal manifestado por el grupo de mujeres, pues precisamente los seis pares en los que la vocal resultante de

la fusión es más próxima a la señal visual aparecen en este grupo, mientras que solo hay dos casos (u / e) = /a/, y (u / e) = /i/, en el grupo de los hombres.

Por último, de todos los casos analizados resulta difícil entender sin embargo la elección que sendos hablantes hacen de la fusión de (o / a) en /i/; de (u / i) en /a/; y de (e / a) en /u/. Como puede verse, los tres pares muestran un comportamiento semejante, ya que la fusión resultante es siempre la vocal más alejada de cada par en el espacio vocálico. No hay que descartar por tanto ejemplos en los que el hablante manifiesta un rechazo total a los estímulos que le llegan por las dos señales.

Fusiones en /a/	(i / o); (o / e); (u / e); (u / i).
Fusiones en /e/	(a / i); (i / a); (a / o); (i / u); (u / a); (u / i).
Fusiones en /i/	(o / a); (u / e).
Fusiones en /o/	(a / u); (u / a); (a / e); (e / u); (u / e); (u / i).
Fusiones en /u/	(e / a).

Tabla 11. *Relación completa de todos los casos de fusión por vocal.*

7. Las mujeres muestran más casos de fusión que los hombres (29 casos en las mujeres frente a solo 9 en los hombres), siendo esta diferencia estadísticamente significativa (χ^2_1=10,66; p=0,001). Asimismo, como se observa en la tabla 7, el grupo de mujeres ofrece también una mayor variabilidad en sus respuestas, especialmente con la *vocal auditiva u*. Estos resultados parecen guardar de nuevo una relación directa con la mayor sensibilidad mostrada por este grupo hacia el canal visual. Si observamos los resultados de fusiones por vocal y sexo (ver tabla 12) resulta particularmente llamativo cómo el porcentaje de fusiones en la vocal /o/ es del 2,1% en los hombres, frente a un 20,3% en las mujeres. La prueba Chi-cuadrado evidencia que estas diferencias para la vocal /o/ son estadísticamente significativas (χ^2_1=8,33; p=0,004), pero no así para el resto de vocales: /a/ (χ^2_1=0,01; p=0,920); /e/ (χ^2_1=3,03; p=0,082); /i/ (χ^2_1=0,14; p=0,735), y /u/ (χ^2_1=1,35; p=0,246).

		Sexo n(%)	
Vocal	Fusión	Hombre	Mujer
a	Sí	4 (8,3)	5 (7,8)
	No	44 (91,7)	59 (92,2)
e	Sí	2 (4,2)	9 (14,1)
	No	46 (95,8)	55 (85,9)
i	Sí	1 (2,1)	2 (3,1)
	No	47 (97,9)	62 (96,9)
o	Sí	1 (2,1)	13 (20,3)
	No	47 (97,9)	51 (79,7)
u	Sí	1 (2,1)	
	No	47 (97,9)	64 (100,0)

Tabla 12. *Prueba Chi-cuadrado resultado fusión por vocal y sexo.*

4.2. Bloque segundo (serie visual)

En esta segunda parte del experimento se expuso a los 28 jueces a una prueba de percepción de las 5 vocales transmitidas únicamente por el canal visual. Cada una de las vocales estaba dispuesta de forma aleatoria en dos series, contabilizándose por tanto 10 estímulos (280 respuestas, 56 por vocal). Como objetivo principal en este caso se plantea averiguar la existencia de diferencias entre la percepción visual de las vocales, centrándonos en cuáles son las vocales que generan más confusión por este canal. Como objetivo secundario se determinará si algún sexo destaca frente al otro de forma significativa cuando se pone a prueba su correcta percepción de esta señal. El resultado de este experimento puede verse en la figura 10, donde se representa por vocales el porcentaje de acierto sobre el total en cada una de las dos series.

Si se analiza a continuación el gráfico de la figura 11, donde se representa la media del porcentaje de acierto y error por vocal, se puede ver que la /a/ es percibida de forma correcta por el 100% de los jueces. La siguen en porcentajes de reconocimiento la /o/ (98,2%), la /u/ (92,9%), la /i/ (85,7%) y la /e/ (82,1%). Estos

resultados guardan relación directa con el peso que tenía cada vocal en la señal visual cada vez que se señalaba este estímulo en condiciones audiovisuales incongruentes (figura 5), aunque en ese caso el orden era ligeramente diferente (de mayor a menor en porcentaje: /o, a, u, i, e/). Los resultados permiten afirmar de este modo que tanto la /a/ como la /o/ son las vocales del español con mayor influencia en la señal visual en condiciones unimodales y bimodales. El valor del coeficiente de correlación entre ambas clasificaciones mediante Spearman es de r=0,718, muy elevado, y no significativo con p=0,086.

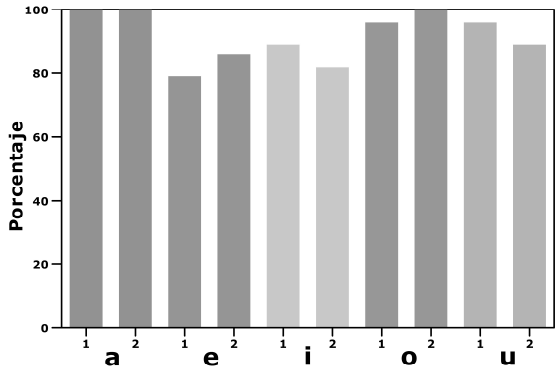


Figura 10. *Porcentaje de acierto visual en cada serie.*

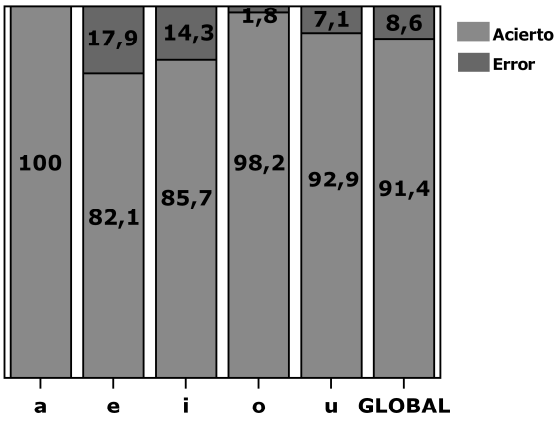


Figura 11. *Media del porcentaje de acierto y error por vocal.*

Por otra parte, como se puede observar en la tabla 13, las vocales que no poseen rasgos vocálicos visualmente relevantes para su identificación en la señal visual (grado máximo de abertura y redondeamiento labial), como son la /e/ y la /i/, son las que presentan los peores porcentajes de acierto, porcentajes que además están situados bastante por debajo del porcentaje de acierto medio global (92%). Considerando estos resultados, la prueba Chi-cuadrado muestra precisamente que el porcentaje de aciertos de las dos vocales con mejores resultados (/a/ = 100% y /o/ = 98,2%) es significativamente superior (χ^2=17,81; p=0,001), al de las dos vocales con peores resultados (/e/ = 82,1%, e /i/= 85,7%). Asimismo, el porcentaje de acierto de la vocal /a/ es significativamente superior al de las vocales /e, i, u/.

	Vocal n (%)				
Resultado	a	e	i	o	u
Acierto	56a (100,0)	46b (82,1)	48b (85,7)	55a, c (98,2)	52b, c (92,9)
Error		10b (17,9)	8b (14,3)	1a, c (1,8)	4b, c (7,1)

Chi-cuadrado: valor=17,81; gl=4; p=0,001

Tabla 13. *Prueba Chi-cuadrado resultado visual por vocal (condiciones unimodales visuales).*

Con el fin de analizar el tipo de vocales que han señalado erróneamente los jueces en todos los casos, se ha elaborado una tabla cruzada que permite observar la cantidad y calidad de las respuestas incorrectas (ver tabla 14). A la vista de esta tabla, parece evidenciarse que los errores se concentran dentro del grupo de vocales posteriores y vocales anteriores, especialmente en este último. De este modo, resulta llamativo el caso de los jueces que erróneamente han seleccionado en hasta 10 ocasiones (un 17,85%) la /i/ cuando era la /e/ la vocal cuya pronunciación visualizaban. Ocurre lo mismo a la inversa, donde hasta en ocho veces se ha seleccionado la /e/ cuando la vocal que realmente se articulaba se correspondía con la /i/ (un 14,28%). En el caso de las vocales posteriores, es la /u/ la que más confusión genera, pues su pronunciación se confunde en hasta cuatro ocasiones con la de la /o/ (7,14%).

	Vocal seleccionada erróneamente *(entre paréntesis el número de casos)*					
Vocal Visual		a	e	i	o	u
	a		0	0	0	0
	e	0		17, 85% (10)	0	0
	i	0	14,28 % (8)		0	0
	o	1,78 % (1)	0	0		0
	u	0	0	0	7,14% (4)	

Tabla 14. *Relación de vocales seleccionadas erróneamente.*

En la figura 12 se representan a continuación las respuestas diferenciadas por sexo ante el estímulo visual, a partir de la media de las dos series presentadas. Como se aprecia, la identificación certera del estímulo (100%) se produce de un modo desigual entre hombres y mujeres, de manera que: (1) el único caso de plena coincidencia se produce como vimos ante el estímulo /a/, cuyo acierto es del 100% en ambos grupos; (2) por su parte la respuesta ante el estímulo visual /o/ produce un 100% de acierto en las dos series para las mujeres, frente a la respuesta /a/ en una de las dos series para los hombres; (3) en la serie de la /u/, los hombres responden en ambas series con /u/ en un 100% de los casos, mientras que en las dos series de /u/ para las mujeres aparece también como opción la respuesta /o/; (4) las series /i/ y /e/ producen una respuesta muy pareja en ambos sexos en cuanto a la variabilidad de las vocales identificadas: /i, e/ y /e, i/. Sin embargo conviene señalar que en las dos series de ambas vocales, para el grupo de hombres la serie primera y segunda (ii), tanto para /i/ como para /e/ resultan idénticas, esto es, ambas series son un espejo, ofreciendo por tanto en todos los casos el mismo número y proporción de respuestas «diferentes», cosa que no sucede con las mujeres en ninguno de los cuatro casos: *i, i (ii), e, e (ii).*

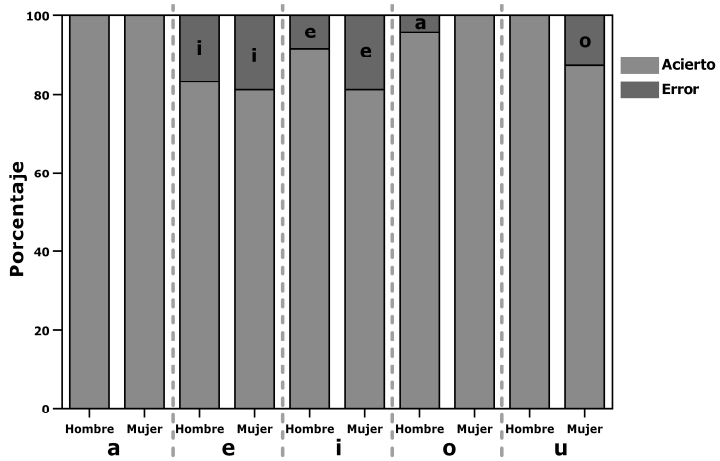


Figura 12. *Relación de respuestas de hombres y mujeres ante el estímulo visual.*

Para terminar con las conclusiones de este apartado, se atiende ahora al objetivo secundario, que intenta determinar si existe algún grupo (se espera que sea el femenino, al que como vimos se le atribuye una mejor capacidad visual) que destaque frente al otro. Sin embargo, y sorprendentemente, es el grupo de los hombres el que goza de un mayor porcentaje de acierto visual en estas condiciones, concretamente un 94,1% de los hombres frente a un 90,6% de las mujeres, aunque estas diferencias no son estadísticamente significativas (χ^2_1=2,17; p=0,1407). Si localizamos ahora las diferencias para cada una de las vocales en cada uno de los grupos, y su relación con la media (figura 13), observamos cómo el porcentaje de aciertos de los jueces masculinos destaca, siendo superior al global, en las vocales /i/ y /u/. En cuanto a ellas, es únicamente la /o/ la vocal cuyo porcentaje de aciertos destaca sobre los jueces masculinos. La prueba Chi-cuadrado (ver tabla 15) evidencia también que las diferencias de aciertos entre las vocales no es estadísticamente significativa: /e/ (χ^2_1=0,41; p=0,084); /i/ (χ^2_1=1,22; p=0,270); /o/ (χ^2_1=1,36; p=0,244), y /u/ (χ^2_1=3,23; p=0,072).

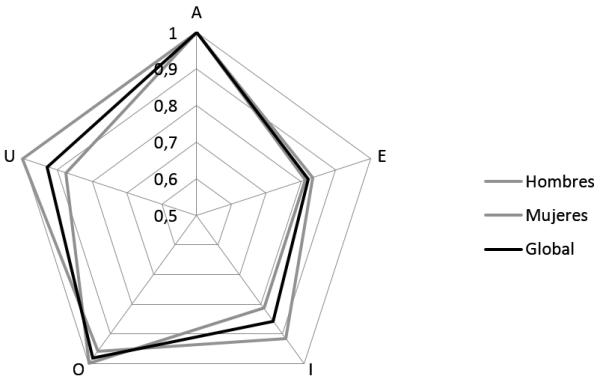


Figura 13. *Porcentaje de acierto visual en condiciones unimodales por vocal y sexo.*

		Sexo n(%)	
Vocal	Resultado	Hombre	Mujer
a	Acierto	24 (100,0)	32 (100,0)
	Error		
e	Acierto	20 (83,3)	26 (81,3)
	Error	4 (16,7)	6 (18,8)
i	Acierto	22 (91,7)	26 (81,3)
	Error	2 (8,3)	6 (18,8)
o	Acierto	23 (95,8)	32 (100,0)
	Error	1 (4,2)	
u	Acierto	24 (100,0)	28 (87,5)
	Error	0 (0,0)	4 (12,5)

Tabla 15. *Prueba Chi-cuadrado resultado visual (condiciones unimodales) por vocal y sexo.*

4.3. Bloque tercero (serie auditiva)

En esta tercera parte del experimento los jueces tuvieron que realizar una prueba de percepción de las 5 vocales transmitidas únicamente por el canal auditivo. Como se esperaba, dada la simplicidad del sistema vocálico español, la identificación certera del estímulo (100%) se produce de forma absoluta y por igual tanto en hombres como en mujeres en todos los casos, por lo que no pueden considerarse diferencias por sexo. Este resultado tan alto permite confirmar por otra parte la validez de este tipo de estímulo auditivo. A la vista de los resultados queda patente cómo la percepción vocálica en español de la señal auditiva en condiciones unimodales resulta mucho más robusta en todas las vocales (a excepción de /a/, que es reconocida por el 100% en las dos modalidades) que la de la señal visual. Esta diferencia entre el porcentaje de acierto de las vocales de la señal visual (92%) y auditiva (100%) en condiciones unimodales es estadísticamente significativa ($\chi^2_1=10{,}23;\ p=0{,}001$).

4.4. Bloque cuarto (*entre la certeza y la duda)*

Como se señaló anteriormente, para la elaboración y ejecución del test se ha querido medir también el grado de certeza en la percepción de los estímulos. De este modo se presentaban en las tres series un apartado de tres opciones que permitía a los sujetos evaluar si habían dudado (y medirlo de algún modo) ante la respuesta seleccionada. Las tres posibles respuestas a la pregunta «¿Estás seguro/a?» eran: «sí / quizá / no».

A continuación se recoge el resumen del conteo de respuestas, así como algunas reflexiones de tipo general sobre las hipótesis y tendencias que pueden verse:

4.4.1. Serie primera (estímulos audiovisuales)

(1) El porcentaje de respuestas «sí» (grado máximo de seguridad) es generalmente mayor cuando los estímulos auditivos y visuales son los mismos. De este modo, podemos encontrar un 100% de respuestas «sí» en los estímulos a/a; o/o y e/e tanto para hombres como para mujeres (y *visual i / auditivo e* para hombres). En los restantes pares con idéntico estímulo visual y auditivo (*i/i* y *u/u*), se producen porcentajes de duda en torno al 6% para ambos sexos, y ello de modo independiente a que se haya producido en el caso de la *u/u* en los hombres un 100% de respuestas correctas.

(2) Si se toma como referencia la vocal auditiva del par audiovisual y se analizan los tamaños de los porcentajes de duda (es decir, aquellos que reúnen las respuestas «no» y «quizás») y de no duda por vocal (ver tabla 15), se observa que los mayores porcentajes de duda están, por este orden, en las vocales /u, o, i, a, e/. Estos porcentajes de duda guardan una estrecha relación con los porcentajes de error auditivo de las vocales que se observaron en la figura 5, ya que estas vocales aparecen también dispuestas en el mismo orden en función de un tamaño del error distribuido de mayor a menor. El valor del coeficiente de correlación entre ambas clasificaciones mediante Spearman es por tanto de 1 (r=1). En cuanto a la relación que mantienen las vocales entre sí en función del grado de certeza (ver tabla 16), la prueba Chi-cuadrado evidencia que el porcentaje de duda de la vocal /u/ con respecto a /a, e, i/, es mayor de un modo estadísticamente significativo (p<0,001). Asimismo, la vocal con menor porcentaje de duda (/e/) presenta unas diferencias estadísticamente significativas con respecto a /i, o, u/.

	Vocal n (%)				
Resultado	a	e	i	o	u
No duda	103a, b (92,0)	105b (93,8)	96a (85,7)	94a, c (83,9)	82c (73,2)
Duda	9a, b (8,0)	7b (6,3)	16a (14,3)	18a, c (16,1)	30c (26,8)

Chi-cuadrado: Valor=24,06; gl=4; p<0,001

Tabla 16. *Prueba Chi-cuadrado resultado certeza por vocal auditiva en condiciones audiovisuales.*

(3) Si se toma ahora como referencia la vocal visual del par audiovisual y se analizan los tamaños de los porcentajes de duda y de no duda por vocal (ver tabla 16), se observa que los mayores porcentajes de duda están, por este orden, en las vocales /a, o, i, e, u/. Esto no parece corresponderse con los resultados de error visual obtenidos para estas vocales en estas condiciones (ver figura 6), ya que estas vocales presentan el siguiente orden de mayor a menor error: /i, e, a-u, o/ (la /a/ y la /u/ reflejan el mismo valor). Se calculó de nuevo la correlación de ambas clasificaciones mediante Spearman, y se obtuvo una correlación negativa (r= -0,105) y no significativa con p= 0,400, de modo que apenas presentan sentido de covariación. En cuanto a la relación que mantienen las vocales entre sí en función

del grado de certeza (ver tabla 17), la prueba Chi-cuadrado muestra que no existen diferencias entre ellas de un modo estadísticamente significativo, con p=0,196.

	Vocal n (%)				
Resultado	a	e	i	o	u
No duda	91a (81,3)	98a (87,5)	96a (85,7)	94a (83,9)	103a (92,0)
Duda	21a (18,8)	14a (12,5)	16a (14,3)	18a (16,1)	9a (8,0)

Chi-cuadrado: Valor=6,05; gl=4; p=0,196

Tabla 17. *Prueba Chi-cuadrado resultado certeza por vocal visual en condiciones audiovisuales.*

(4) El grupo de mujeres muestra más duda que los hombres. El hecho de que este grupo muestre más influencia hacia la señal visual, tal como se vio en el punto 4.1, parece guardar una relación directa con la seguridad de su respuesta, la cual disminuye cuando ambas señales se presentan de forma incongruente. Así, las mujeres presentan un 10,25% de respuestas con «quizá» (frente al 8,3% de los hombres), y un 3% con «no» (frente al 1,6% de los hombres). Incluyendo el «no» y el «quizá», el porcentaje total de duda en las mujeres es mayor que el de los hombres (13,25% frente a 10%), sin que esta diferencia sea estadísticamente significativa / (χ^2_1=2,01; p=0,1562). Por otra parte, si observamos el porcentaje del grado de certeza para cada una de las vocales según el sexo, se puede observar que tanto si tenemos como referencia el estímulo auditivo como el estímulo visual, no se presentan diferencias que sean estadísticamente significativas entre hombres y mujeres (ver tabla 18). Con referencia al estímulo auditivo los valores son: /a/ (χ^2_1=1,70; p=0,192); /e/ (χ^2_1=0,62; p=0,430); /i/ ($\chi^2_1$0,20; p=0,640); /o/ (χ^2_1=0,79; p=0,373); y /u/ (χ^2_1=0,01; p=0,951); y con referencia al estímulo visual: /a/ (χ^2_1=0,96; p=0,328); /e/ (χ^2_1=3,00; p=0,083); /i/ (χ^2_1=0,01; p=0,938); /o/ (χ^2_1=0,14; p=0,710); y /u/ (χ^2_1=0,36; p=0,547).

		Sexo n (%)			
Vocal	Resultado	Hombre	Mujer	Hombre	Mujer
a	No duda	46 (95,8)	57 (89,1)	41 (85,4)	50 (78,1)
	Duda	2 (4,2)	7 (10,9)	7 (14,6)	14 (21,9)
e	No duda	46 (95,8)	59 (92,2)	45 (93,8)	53 (82,8)
	Duda	2 (4,2)	5 (7,8)	3 (6,3)	11 (17,2)
i	No duda	42 (87,5)	54 (84,4)	41 (85,4)	55 (85,9)
	Duda	6 (12,5)	10 (15,6)	7 (14,6)	9 (14,1)
o	No duda	42 (87,5)	52 (81,3)	41 (85,4)	53 (82,8)
	Duda	6 (12,5)	12 (18,8)	7 (14,6)	11 (17,2)
u	No duda	35 (72,9)	47 (73,4)	45 (93,8)	58 (90,6)
	Duda	13 (27,1)	17 (26,6)	3 (6,3)	6 (9,4)
		Con referencia al estímulo auditivo		Con referencia al estímulo visual	

Tabla 18. *Prueba Chi-cuadrado resultado certeza por vocal y sexo en estímulos audiovisuales.*

4.4.2. Serie segunda (estímulos visuales)

(1) Si se analizan los tamaños de los porcentajes de duda y de no duda en la respuesta por vocal de la serie visual (ver tabla 19), se observa que los mayores porcentajes de duda están, por este orden, en las vocales /e, i, u, o, a/. Este orden de las vocales guarda una estrecha relación con el tamaño del error en la identificación, ya que como se pudo observar en la figura 11, las vocales aparecen distribuidas también con el mismo rango en función de un tamaño del error distribuido de mayor a menor. El valor del coeficiente de correlación entre ambas clasificaciones mediante Spearman es por tanto de 1 (r=1). En cuanto a la relación que mantienen las vocales entre sí en función del grado de certeza (ver tabla 19), la

prueba Chi-cuadrado evidencia que el porcentaje de duda de la vocal /a/ (la vocal que presenta menores porcentajes de duda) con respecto a /e, i/ (las vocales que presentan mayores porcentajes) es menor de un modo estadísticamente significativo (p<0,001).

	Vocal n (%)				
Resultado	a	e	i	o	u
No duda	53a (94,6)	33b (58,9)	41b (73,2)	46a, b (82,1)	43a, b (76,8)
Duda	3a (5,4)	23b (41,1)	15b (26,8)	10a, b (17,9)	13a, b (23,2)

Chi-cuadrado: Valor=21,55; gl=4; p<0,001

Tabla 19. *Prueba Chi-cuadrado resultado certeza por vocal visual en condiciones unimodales.*

(2) Si representamos los porcentajes de inseguridad de cada grupo en la respuesta por vocal (esto es, el porcentaje de respuestas donde los jueces de cada grupo señalaron que «no» estaban seguros o que «quizá» no estaban seguros de que fuera esa vocal, ver figura 14), se puede observar que los hombres de la muestra, a pesar de tener un mayor porcentaje de acierto visual en condiciones unimodales, también presentan una mayor inseguridad en todas y cada una de las vocales, lo que hace suponer que su respuesta es más espontánea que la de las mujeres. Es decir: los hombres alcanzan mejores porcentajes de identificación (aunque esta diferencia como se vio en el comentario de la figura 14 no sea estadísticamente significativa), pero dudan más (en concreto un 30,8% de duda frente al 16,8% de las mujeres). El grupo de mujeres en cambio reconoce en términos porcentuales algo menos esas mismas vocales, pero está mucho más seguro de haberlas identificado correctamente. Esta diferencia en el grado de certeza sí es estadísticamente significativa (χ^2_1=5,91; p=0,015).

En la tabla 20 se observa de un modo más detallado el porcentaje del grado de certeza de la serie visual para cada una de las vocales según el tipo de jueces. Para las vocales /a/ y /u/, el porcentaje de no duda en las mujeres es mayor que en los hombres de un modo estadísticamente significativo: /a/ (χ^2_1=4,27; p=0,040) y /u/ (χ^2_1=4,81; p=0,028). En el resto de vocales no se encuentran diferencias

significativas en el grado de certeza entre hombres y mujeres: /e/ (χ^2_1=1,38; p=0,240); /i/ (χ^2_1=0,12; p=0,728); y /o/ (χ^2_1=1,46; p=0,227).

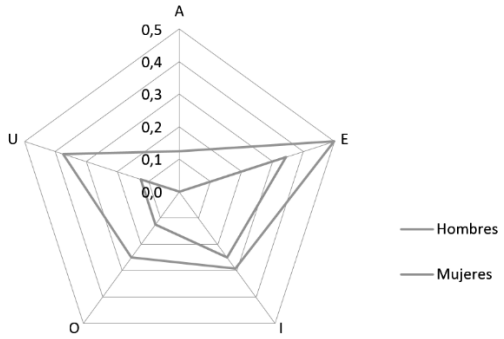


Figura 14. *Porcentaje de inseguridad en la respuesta por vocal y sexo.*

		Sexo n(%)	
Vocal	Resultado	Hombre	Mujer
a	Acierto	21 (87,5)	32 (100,0)
	Error	3 (12,5)	
e	Acierto	12 (50,0)	21 (65,6)
	Error	12 (50,0)	11 (34,4)
i	Acierto	17 (70,8)	24 (75,0)
	Error	7 (29,2)	8 (25,0)
o	Acierto	18 (75,0)	28 (87,5)
	Error	6 (25,0)	4 (12,5)
u	Acierto	15 (62,5)	28 (87,5)
	Error	9 (37,5)	4 (12,5)

Tabla 20. *Porcentaje de inseguridad en la respuesta por vocal y sexo.*

(3) Si se atiende ahora al resumen completo de los datos en función del grado de certeza y acierto por vocal en cada grupo de jueces (ver tabla 21) puede observarse que las vocales que ofrecen un mayor número de dudas en la respuesta serían en primer lugar la /e/ (donde se acumulan la mayoría de los «no», y donde se concentran altos porcentajes de duda en la opción «quizá» para respuestas correctas tanto en hombres como en mujeres), y en segundo lugar la /u/, en la que el porcentaje de duda alcanza el valor más elevado de toda la serie para el «quizá», a pesar de que no aparece el «no» entre las respuestas.

		Grado de certeza con identificación correcta			Grado de certeza con identificación incorrecta		
		Sí	Quizá	No	Sí	Quizá	No
a	M	100	-	-	-	-	-
	H	87,5	12,5	-	-	-	-
e	M	50	28,12	-	15,62	3,12	3,12
	H	45.83	25	12,5	4,16	12,5	-
i	M	62,5	18,75	-	12,5	6,25	-
	H	62,5	29,16	-	8,33	-	-
o	M	87.5	12,5	-	-	-	-
	H	75	16,67	4,16	-	-	4,16
u	M	75	12,5	-	12,5	-	-
	H	62,5	37,5	-	-	-	-

Tabla 21. *Porcentaje de respuestas según el grado de certeza y cierto ante el estímulo visual.*

Por otro lado, las dos vocales mejor reconocidas en condiciones unimodales visuales (/a/ con un 100% y /o/ con un 98% -ver figura 13-) son también las que ofrecen menores porcentajes de duda en la respuesta, aunque como se señalaba, los hombres se muestran más dubitativos. Así, las mujeres tienen un 100% de certeza total (sumando las dos series) en la identificación de /a/, mientras que para este mismo estímulo los hombres dudan y apuntan la respuesta «quizá» en ambas series

(8,33% y 16,67%). Para la /o/, a pesar de que en el grupo de mujeres la respuesta en todos los casos es /o/ y por tanto hay un 100% de acierto, en ambas series se produce un 87,50% de «sí» frente a un 12,50% de «quizá». Para este mismo estímulo los hombres se muestran más dubitativos, y solamente tienen acierto total en una de las series. Además, sus porcentajes de duda son más altos, ya que el «quizá» sube a valores de 16,67% en las dos series, y en ambas aparece también la respuesta «no» con un 8,33%.

(4) Si se analizan las vocales en función de los casos en los que los hablantes han señalado que no estaban en absoluto seguros de su identificación, y que por tanto marcaron la respuesta «no» (ver tabla 22), se observa que solo la /e/ muestra diferencias estadísticamente significativas con respecto al resto de vocales (p=0,028).

	Vocal n (%)				
Resultado	a	e	i	o	u
No duda	56a (100,0)	52b (92,9)	56a (100,0)	54a (96,4)	56a (100,0)
Duda		4b (7,1)		2a (3,6)	

Chi-cuadrado: Valor=10,90; gl=4; p=0,028

Tabla 22. *Prueba Chi-cuadrado resultado inseguridad total por vocal visual en condiciones unimodales.*

Finalmente, si se atiende a las diferencias en el grado de certeza entre hombres y mujeres en función de las respuestas producidas con «no», se aprecia que el porcentaje más alto de presencia del «no» aparece en hombres (4,1% frente al 0,83% de las mujeres), sin que estas diferencias puedan considerarse significativas ($\chi^2_1=1,54$; p=0,214). Esta falta de significatividad estadística se obtiene también si se analizan estas diferencias vocal por vocal (ver tabla 23), donde ni la /o/ ($\chi^2_1=1,82$; p=0,178), ni la /e/ ($\chi^2_1=2,77$; p=0,0967) presentan diferencias significativas entre hombres y mujeres.

		Sexo n(%)	
Vocal	Resultado	Hombre	Mujer
a	No		
	Otras	24 (100,0)	32 (100,0)
e	No	3 (12,5)	1 (3,1)
	Otras	21 (24)	31 (96,9)
i	No		
	Otras	24 (100,0)	32 (100)
o	No	2 (8,3)	
	Otras	22 (91,7)	32 (100)
u	No		
	Otras	24 (100,0)	32 (100,0)

Tabla 23. *Porcentaje de inseguridad total en la respuesta por vocal y sexo.*

4.4.3. Serie tercera (estímulos auditivos)

1. Se producen respuestas «sí» (total seguridad) tanto para hombres como mujeres en un 100% de los casos ante los estímulos /i, a, o/.

2. Se produce una respuesta de «sí» para el 100% de las mujeres en el caso de la /e/, pero en los hombres aparece un 8,33% de duda, a través de la respuesta «quizá».

3. Se produce una respuesta de «sí» para el 100% de los hombres en el caso de la /u/, pero en las mujeres aparece un 6,25% de duda con la respuesta «quizá».

4. En ningún caso, y en ningún sexo, se produce la respuesta «no».

5. CONCLUSIONES

Los resultados de este experimento permiten extraer varias conclusiones. En primer lugar, en los casos de percepción audiovisual con estímulos discordantes se encontraron diferencias estadísticamente significativas entre el porcentaje de veces que los hablantes se decantaban acertadamente por la señal auditiva sobre la visual, siendo ampliamente mayor el primero (81% frente a 12%). Esto se demostró ser así para todas las vocales y para los dos grupos de hombres y mujeres. Los resultados muestran por tanto que la señal auditiva, con este tipo de estímulos y condiciones para el español, predomina sobre la visual, lo que permite confirmar que la selección opera eligiendo la respuesta del canal que ofrece una información más sólida. Además, tal y como aparece de forma recurrente en la bibliografía, se produjeron también casos de fusión de los estímulos, de modo que en un 7% de los casos los hablantes percibieron una vocal que no procedía de ninguna de estas dos señales. Estos resultados permiten mostrar que, aunque pequeño, dada la fuerte prevalencia de la señal auditiva, en español se produce también un efecto McGurk, ya que en un 19% de los casos los oyentes eligieron una vocal diferente de aquella que habían oído, por interferencia de la señal visual. Estas diferencias entre aciertos y fallos en la percepción del estímulo auditivo son estadísticamente significativas.

En español, el grado de abertura máximo y la presencia de redondeamiento labial son los rasgos vocálicos que ejercen mayor influencia, a nivel visual, en la percepción del estímulo vocálico que llega por esta señal. De este modo /a, o, u/ determinan grados de variabilidad más altos ante claves audiovisuales discordantes, y son, en este orden, las vocales mejor identificadas cuando el estímulo es únicamente visual (en esas condiciones unimodales, /a/ alcanza un 100% en todos los casos). En este experimento se vio también que los estímulos auditivos /o, u/ transmitidos en condiciones audiovisuales discordantes obtenían peores porcentajes de acierto auditivo que /a, e, i/, siendo estas diferencias estadísticamente significativas. Se cree que aunque todas las vocales fueron identificadas por el 100% de los hablantes en condiciones unimodales auditivas, y sus valores acústicos están dentro de los campos de dispersión vocálica señalados por la bibliografía (Martínez Celdrán y Fernández Planas, 2007), los menores porcentajes de acierto auditivo se deben al marcado redondeamiento labial (con vistas a producir en este experimento un gesto articulatorio visualmente bien definido) y a su influencia para la configuración de unos formantes (especialmente F2) ligeramente más bajos, lo que a la postre redunda en la configuración de unos timbres particularmente más oscuros.

En los casos de fusión a partir de señales audiovisuales incongruentes, se comprobó que la vocal percibida tiende a situarse en un espacio vocálico contiguo al de la vocal auditiva, y/o a situarse en una posición equidistante entre las vocales que proceden de las dos señales. Dada la simplicidad del sistema vocálico español, los casos de fusión se concentraron en las vocales situadas en las partes medias y bajas del espacio vocálico, como son /a, e, o/. De este modo, cualquier combinación audiovisual de una vocal cerrada /i, u/ con /a/ (los extremos del «triángulo» vocálico) produce como único caso de fusión la vocal semicerrada más equidistante de las dos. Por ejemplo, los pares (i / a; a / i) dieron /e/, y los pares (u / a; a / u) dieron /o/ (aunque el par (*u / a)* da también /e/). Considerado también el reducido número de segmentos que pueden intervenir en la fusión, se encuentra también una mayoría de pares «espejo» que, independientemente del cruce entre estímulos visuales o auditivos, producen un mismo caso de fusión, como (a / i; i / a), (a / u; u / a); (e / u; u / e); (i / u; u / i), que representan el 40% de las combinaciones audiovisuales discordantes posibles, y el 53,3% de las combinaciones en las que se producen estos casos de fusión. Por otra parte, cuando se presentan cruzados los pares audiovisuales (e / i) y (o / u), el estímulo percibido siempre procede de una de estas dos señales, y nunca se producen casos de fusión.

Uno de los objetivos de este experimento se centraba también en determinar la existencia de diferencias en la percepción de estos estímulos entre hombres y mujeres. Así, se observó que en este tipo de condiciones audiovisuales discordantes existen diferencias estadísticamente significativas en cuanto al porcentaje de acierto auditivo y acierto visual. Aunque como vimos ambos grupos se decantan por la información que llega de la señal auditiva, se comprueba por un lado que los hombres presentan un mayor nivel de acierto en el estímulo auditivo, y por otro que las mujeres muestran una mayor preferencia que los hombres por el estímulo visual, así como una mayor variabilidad en la respuesta. Esto se corresponde con los estudios que demuestran una mayor sensibilidad de las mujeres sobre los hombres ante las señales visuales (Aloufy *et al*, 1996; Bayliss *et al*, 2005; Traunmüller y Öhrström, 2007).

De este modo, como puede apreciarse en la tabla 5 de este estudio, el grupo de mujeres es el único que en al menos uno de los estímulos audiovisualmente incongruentes opta por la información visual sobre la auditiva: esto se produce en el estímulo *visual o /auditivo u*. En otros dos casos, como en el estímulo *visual a / auditivo u*, y en *visual u / auditivo o* se presentan respectivamente valores por encima del 30% e incluso del 40%. En los hombres en cambio prevalece siempre lo auditivo, y solamente hay un único par (*visual o /auditivo u*) que llegue a alcanzar los valores del 30%. En el resto de pares, y en la línea de los resultados

anteriores, las mujeres tienden a alejarse también más de la selección auditiva. Como se dijo también, las mujeres ofrecen una mayor variabilidad de respuestas posibles frente a los mismos estímulos que los hombres. Así, en 10 de los 25 estímulos las mujeres registran más respuestas posibles que el grupo masculino, hay 13 ocasiones de 25 donde hombres y mujeres ofrecen igual número de variabilidad en la respuesta, y solo hay dos ejemplos en los que los hombres registran una mayor variabilidad.

La mayor influencia que la señal visual ejerce sobre el grupo de mujeres en condiciones audiovisuales incongruentes no parece ser una ventaja en la lectura labiofacial (esto es, cuando los estímulos se presentan en condiciones unimodales visuales) tal y como se hubiera esperado, ya que obtienen peores datos de identificación que los hombres. Sin embargo, estas diferencias no tienen significación estadística. Sí se observan en cambio diferencias estadísticamente significativas entre los niveles de certeza de uno y otro grupo, pues las mujeres, aun obteniendo ligeramente peores resultados en la identificación, se muestran mucho más seguras de haber identificado la vocal correcta (concretamente un 83,2% de sus respuestas se emiten con un grado de seguridad total, frente al 69,2% de los hombres). En relación con las diferencias porcentuales de identificación vocálica entre hombres y mujeres en estas condiciones, el grupo de los hombres identifica sin errores las series /a/ y /u/, y las mujeres las series /a/ y /o/. En cambio la vocal /e/ ofrece los peores resultados de identificación cuando el estímulo es solo visual, tanto para hombres como para mujeres.

Se demuestra con este experimento por tanto que, aun predominando con fuerza el canal auditivo, la información audiovisual en la percepción vocálica del español es complementaria, y que la presencia de estímulos visuales incongruentes afecta su propia percepción, aunque la información visual por sí sola no sea suficiente para la discriminación perfecta entre todas las vocales. Dicho esto, es evidente que el español es una lengua con un reducido inventario vocálico que no presenta problemas de discriminación auditiva entre vocales en condiciones unimodales, y que el redondeamiento labial, una de las marcas visuales más informativas de estos segmentos, no tiene valor fonológico en nuestra lengua. Sin embargo, resulta no menos cierto que nuestras vocales también se ven, y que los desajustes en la señal audiovisual pueden dar lugar (dan lugar) a confusiones, aunque su número sea mucho más reducido en relación con otras lenguas. Es preciso decir también que el pequeño tamaño de la muestra considerada en este experimento nos obliga a ser prudentes a la hora de realizar afirmaciones categóricas y extrapolables a la población general, y que sería deseable considerar en futuros estudios muestras más amplias para realizar análisis inferenciales. En este sentido, se abren también

futuras líneas de investigación en análisis que consideren aumentar el número de informantes (lo que incluye como mínimo un hombre y una mujer), incluir muestras extraídas de distintos estilos de habla, de distintos contextos (considerando márgenes consonánticos simétricos y asimétricos), de distintas posiciones silábicas dentro de la palabra, y de distintas posiciones acentuales. Se pueden observar también los posibles cambios en función del grupo de edad (McGurk y McDonald, 1976; Rosenblum *et al*, 1997), o el efecto producido en hablantes que tienen el español como L2 y poseen distintos inventarios vocálicos (Reisberg et al, 1987). Considerados estos aspectos, el propósito del presente estudio ha sido ofrecer nuevos datos sobre la relevancia de los canales en los procesos de percepción audiovisual del habla para esta clase de estímulos y condiciones, observar cómo se comporta el efecto McGurk, señalado en otras lenguas, y proponer posibles diferencias de sensibilidad hacia la señal visual en función del sexo del hablante.

6. REFERENCIAS BIBLIOGRÁFICAS

ABELIN, Å. (2007): «Emotional McGurk effect in Swedish», en L. Berthouze, C. G. Prince, M. Littman, H. Kozima y C. Balkenius (eds.): *Proceedings of the Seventh International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund University Cognitive Studies, 135, pp. 73-76.

ALOUFY, S.; M. LAPIDOT y M. MYSLOBODSKY (1996): «Differences in susceptibility to the "blending illusion" among native Hebrew and English speakers», *Brain and Language*, 53, pp. 51-57.

ARGYLE, M. y R. INGHAM (1972): «Gaze, mutual gaze, and proximity», *Semiotica*, 6, pp. 32-49.

ASSMANN, P. F. y A. Q. SUMMERFIELD (2004): «The perception of speech under adverse conditions», en S. Greenberg, W. A. Ainsworth, A. N. Popper y R. R. Fay (eds.): *Processing in the auditory system*, Heidelberg, Springer-Verlag, pp. 231-308.

BAYLISS, A. P.; G. DI PELLEGRINO y S. P. TIPPER (2005): «Sex differences in eye gaze and symbolic cueing of attention», *The Quarterly Journal of Experimental Psychology*, 58A (4), pp. 631–650.

CALVERT, G. A.; E. T. BULLMORE, M. J. BRAMMER, R. CAMPBELL, S. C. R. WILLIAMS, P. K. MCGUIRE, P. W. R. WOODRUFF, S. D. IVERSEN y A. S. DAVID (1997): «Activation of Auditory Cortex During Silent Lipreading», *Science*, 276, pp. 593-596.

CAMPBELL, R. (1994): «Audiovisual Speech: Where, what, when, how?», *Current Psychology of Cognition*, 13, pp. 76-80.

CHEN, Y. y V. HAZAN (2007): «Language effects on the degree of visual influence in audiovisual perception», en J. Trouvain y W. J. Barry (eds.): *Proceedings of the 16th International Congress of the Phonetic Sciences*, Saarbrücken, Universidad de Sarland, pp. 2177-2180.

COLIN, C.; M. RADEAU y P. DELTENRE (2005): «Top-down and bottom-up modulation of audiovisual integration in speech», *European Journal of Cognitive Psychology*, 17 (4), pp. 541- 560.

DODD, B. (1977): «The role of vision in the perception of speech», *Perception*, 6, pp. 31-40.

DODD, B. (1979): «Lip-reading in infants: attention to speech presented in- and out-of-synchrony», *Cognitive Psychology*, 11 (4), pp. 478-84.

DARWIN, C. J. y R. P. CARLYON (1995): «Auditory grouping», en B. C. J. Moore (ed.): *The handbook of perception and cognition*, Londres, Academic Press, pp. 387-424.

ERBER, N. P. (1969): «Interaction of audition and vision in the recognition of oral speech stimuli», *Journal of Speech and Hearing Research*, 12, pp. 423-425.

FOWLER, C. A. y D. J. DEKLE (1991): «Listening with eye and hand: Cross-Modal contributions to speech perception», *Journal of Experimental Psychology: Human Perception and Performance*, 17 (3), pp. 816-828.

GAIL S. D.; K. T. ELIZABETH y L. R. CATHERINE (2010): «Vowel identification by younger and older listeners: Relative effectiveness of vowel edges and vowel centers», *The Journal of Acoustical Society of America*, 128 (3), pp. 105-110.

GICK, B. y D. DERRICK (2009): «Aero-tactile integration in speech perception», *Nature*, 462, pp. 502-504.

GIL FERNÁNDEZ, J. (2007): *Fonética para profesores de español: de la teoría a la práctica*, Madrid, Arco Libros.

GIRIN L.; J. L. SCHWARTZ y G. FENG (2001): «Audio-Visual Enhancement of Speech in Noise», *The Journal of Acoustical Society of America*, 109 (6), pp. 3007-3020.

GREEN, K. P. y A. GERDEMAN (1995): «Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels», *Journal of Experimental Psychology: Human Perception and Performance*, 21 (6), pp. 1409-1426.

GREEN, K. P. y P. K. KUHL (1989): «The role of visual information in the processing of place and manner features in speech perception», *Perception & Psychophysics*, 45 (1), pp. 34-42.

GREEN, K. P.; P. K. KUHL, A. N. MELTZOFF y E. B. STEVENS (1991): «Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect», *Perception & Psychophysics*, 50 (6), pp. 524-536.

GREEN, K. P. y L. W. NORRIX (1997): «Acoustic cues to place of articulation and the McGurk Effect: The role of release bursts, aspiration, and formant transitions», *Journal of Speech, Language, and Hearing Research*, 40 (3), pp. 646-665.

IRWIN, J. R.; D. H. WHALEN y C. A. FOWLER (2006): «A sex difference in visual influence on heard speech», *Perception & Psychophysics*, 68 (4), pp. 582-592.

JENKINS, J. J.; W. STRANGE y T. R. EDMAN (1983): «Identification of vowels in 'vowelless' syllables», *Perception & Psychophysics*, 34 (5), pp. 441-450.

JOHNSON, F. M.; L. HICKS, T. GOLDBERG y M. MYSLOBODSKY (1988): «Sex differences in Lipreading», *Bulletin of the Psychonomic Society*, 26 (2), pp. 106-108.

JONES, D. (1917): *An English Pronouncing Dictionary*, Londres, Dent.

KANZAKI, R. y R. CAMPBELL (1999): «Effect of facial brightness reversal on visual and audiovisual speech perception», en D. Massaro (ed.): *Audio Visual*

Speech Processing International Conference, University of California, Santa Cruz.

KEWLEY-PORT, D.; Z. T. BURKLE y J. H. LEE (2007): «Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing impaired listeners», *The Journal of the Acoustical Society of America*, 122 (4), pp. 2365-2375.

KIM, J. y C. DAVIS (2003): «Hearing foreign voices: does knowing what is said affect masked visual speech detection?», *Perception*, 32 (1), pp. 111–120.

KUHL P. K. y A. MELTZOFF (1982): «The bimodal perception of speech in infancy», *Science*, 218, pp. 1138-1141.

LADEFOGED, P. (2001): *Vowels and consonants: an introduction to the sounds of languages*, Malden, Blackwell Publishers Ltd.

LADEFOGED, P. y K. JOHNSON (1975): *A course in phonetics*, Wadsworth, Cengage Learning, 2006[6].

LADEFOGED, P. e I. MADDIESON (1996): *The sounds of the world's languages*, Oxford, Blackwell Publishers Ltd.

LEEB, R. T. y F. G. REJSKIND (2004): «Here's looking at you, kid! A longitudinal study of perceived gender differences in mutual gaze behavior in young infants», *Sex Roles*, 50 (1), pp. 1–14.

LISKER, L. y M. ROSSI (1992): «Auditory and visual cueing of the [+/- rounded] feature of vowels», *Language and Speech*, 35 (4), pp. 391-417.

MACDONALD, J. (2006): «Hearing Lips and Seeing Voices: Illusion and Serendipity in Auditory-Visual Perception Research», en J. Atkinson y M. Crove (eds.): *Interdisciplinary Research: Diverse Approaches in Science, Technology, Health and Society*, John Wiley & Sons, Chichester, pp. 101-115.

MACDONALD, J.; S. ANDERSEN y T. BACHMAN (2000): «Hearing by eye: how much spatial degradation can be tolerated?», *Perception*, 29 (10), pp. 1155-1168.

MACLEOD, A. y Q. SUMMERFIELD (1987): «Quantifying the contribution of vision to speech perception in noise», *British Journal of Audiology*, 21 (2), pp. 131-141.

MARTÍNEZ CELDRÁN, E. y A. MA. FERNÁNDEZ PLANAS (2007): *Manual de fonética española: articulaciones y sonidos del español*, Barcelona, Ariel Lingüística.

MASSARO, D. W. (1989): «Testing between the TRACE Model and the Fuzzy Logical Model of Speech perception», *Cognitive Psychology*, 21 (3), pp. 398–421.

MASSARO, D. W. (1998): *Perceiving talking faces: From speech perception to a behavioral principle*, Cambridge, Massachusetts, MIT Press.

MASSARO, D. W. y M. M. COHEN (1990): «Perception of synthesized audible and visible speech», *Psychological Science*, 1 (1), pp. 55-63.

MASSARO, D. W. y M. M. COHEN (1996): «Perceiving speech from inverted faces», *Perception and Psychophysics*, 58 (7), pp. 1047-1065.

MASSARO, D. W.; L. A. THOMPSON, B. E. BARRON y E. LAREN (1986): «Developmental changes in visual and auditory contributions to speech perception», *Journal of Experimental Child Psychology*, 41 (1), pp. 93-113.

MCGURK, H. e I. MACDONALD (1976): «Hearing lips and seeing voices», *Nature*, 264, pp. 746-748.

MORAIN, G. G. (2001): «Kinesics and cross-cultural understanding», en J. M. Valdes (ed.): *Culture Bound*, Cambridge University Press, pp. 64-76.

MUNHALL, K. G.; P. GRIBBLE, L. SACCO y M. WARD (1996): «Temporal constraints on the McGurk effect», *Perception and Psychophysics*, 58 (3), pp. 351-362.

MUNHALL, K. G. y Y. TOKHURA (1998): «Audiovisual gating and the time course of speech perception», *The Journal of the Acoustical Society of America*, 104 (1), pp. 530–539.

MURASE, M.; D. N. SAITO, T. KOCHIYAMA, H. C. TANABE, S. TANAKA; T. HARADA, Y. ARAMAKI, M. HONDA y N. SADATO (2008): «Cross-modal integration during vowel identification in audiovisual speech: a functional magnetic resonance imaging study», *Neuroscience letters*, 434 (1), pp. 71-76.

NIELSEN, K. (2004): «Segmental differences in the visual contribution to speech intelligibility», *UCLA Working Papers in Phonetics*, 103, pp. 106-147.

O'SHEA, M. (2005): *The Brain: A Very Short Introduction*, Oxford University Press.

OSTRAND, R.; SH. BLUMSTEIN y J. MORGAN (2011): «When hearing lips and seeing voices becomes perceiving speech: auditory-visual integration in lexical access», en L. Carlson, C. Hölscher y T. Shipley (eds.): *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin, Cognitive Science Society, pp. 1376-1381.

PARÉ, M.; C. RICHLER, M. HOVE y K. MUNHALL (2003): «Gaze behavior in audiovisual speech perception: The influence on ocular fixations on the McGurk effect», *Perception and Psychophysics*, 65 (4), pp. 533-567.

RAHMAWATI, S. y M. OHGISHI (2011): «Cross cultural studies on audiovisual speech processing: the McGurk effects observed in consonant and vowel perception», en T. Juhana, A. Munir Iskandar y N. Rachmana Hendrawan (eds.): *Proceedings of the 6th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2011*, pp. 59-63.

REISBERG, D.; J. MCLEAN y A. GOLDFIELD (1987): «Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli», en B. Dodd y R. Campbell (eds.): *Hearing by Eye: The Psychology of Lip-reading*, Lawrence Erlbaum Associates Ltd, pp. 97-113.

RICHARDSON, A. C. (2010): «Effect of Visual Input on Vowel Production in English Speakers», *Linguistics Honors Projects*, Paper 5. http://digitalcommons.macalester.edu/ling_honors/5 [22/09/2013]

RIZZOLATTI, G. y L. CRAIGHERO (2004). «The Mirror-Neuron System», *Annual Review of Neuroscience*, 27, pp. 169-192.

ROBERT-RIBES, J.; J.-L. SCHWARTZ, T. LALLOUACHE y P. ESCUDIER (1998): «Complementary and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise», *The Journal of Acoustical Society of America*, 103 (6), pp. 3677-3689.

ROSENBLUM, L. D. y H. M. SALDAÑA (1996): «An audiovisual test of kinematic primitives for visual speech perception», *Journal of Experimental Psychology: Human Perception and Performance*, 22 (2), pp. 318-331.

ROSENBLUM, L. D.; M. A. SCHMUCKLER y J. A. JOHNSON (1997): «The McGurk effect in infants», *Attention, Perception & Psychophysics*, 59 (3), pp. 347-357.

ROSS, L. A.; D. SAINT-AMOUR, V. M. LEAVITT, D. C. JAVITT y J. J. FOXE (2007): «Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments», *Cerebral Cortex*, 17 (5), pp. 1147–1153.

ROUGER, J.; B. FRAYSSE, O. DEGUINE y P. BARONE (2008): «McGurk effects in cochlear-implanted deaf subjects», *Brain Research*, 1188, pp. 87–99.

SAMS, M.; R. AULANKO, M. HAMALAINEN, R. HARI, O. V. LOUNASMAA, S.-T. LU y J. SIMOLA (1991): «Seeing speech: visual information from lip movements modifies activity in the human auditory cortex», *Neuroscience Letters*, 127, pp. 141-145.

SCHEFFERS, M. T. M. (1983): *Sifting vowels. Auditory pitch analysis and sound segregation*, tesis doctoral, Universidad de Groningen, Países Bajos.

SEKIYAMA, K. y Y. TOHKURA (1991): «McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility», *The Journal of Acoustical Society of America*, 90 (4), pp. 1797-1805.

SEKIYAMA, K. y Y. TOHKURA (1993): «Inter-language differences in the influence of visual cues in speech perception», *Journal of Phonetics*, 21 (4), pp. 427-444.

SEKIYAMA, K.; D. BURNHAM, H. TAM y D. ERDENER (2003): «Auditory-Visual Speech Perception Development in Japanese and English Speakers», en J.-L. Schwartz, F. Berthommier, M.-A. Cathiard y D. Sodoyer (eds.): *Proceedings of the International Conference on Auditory-Visual Speech Processing*, St. Jorioz, pp. 61-66.

SKIPPER, J. I.; V. VAN WASSENHOVE, H. C. NUSBAUM y S. L. SMALL (2007): «Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception», *Cerebral Cortex*, 17 (10), pp. 2387-2399.

STRANGE, W.; J. J. JENKINS y T. L. JOHNSON (1983): «Dynamic specification of coarticulated vowels», *The Journal of Acoustical Society of America*, 74 (3), pp. 695-705.

STRANGE, W.; R. R. VERBRUGGE, D. P. SHANKWEILER y T. R. EDMAN (1976): «Consonant environment specifies vowel identity», *The Journal of the Acoustical Society of America*, 60 (1), pp. 213-224.

SUMBY, W. H. e I. POLLACK (1954): «Visual contribution to speech intelligibility in noise», *The Journal of the Acoustical Society of America*, 26 (2), pp. 212-215.

SUMMERFIELD, Q. y M. MCGRATH (1984): «Detection and Resolution of Audio-visual Incompatibility in the Perception of Vowels», *Quarterly Journal of Experimental Psychology*, 36A, pp. 51-74.

TRAUNMÜLLER H. y N. ÖHRSTRÖM (2007): «Audiovisual perception of openness and lip rounding in front vowels», *Journal of Phonetics*, 35 (2), pp. 244-258.

TSEVA, R. (1989): «L'arrondissement dans l'identification visuelle des voyelles du français», *Bulletin du Laboratoire de la Communication Parlée*, 3, pp. 149-186.

VALKENIER, B.; J. Y. DUYNE, T. C. ANDRINGA y D. BAŞKEN (2012): «Audiovisual Perception of Congruent and Incongruent Dutch Front Vowels», *Journal of Speech, Language, and Hearing Research*, 55 (6), pp. 1788-1801.

VAN WASSENHOVE, V.; K. W. GRANT y D. POEPPEL (2007): «Temporal window of integration in auditory-visual speech perception», *Neuropsychologia*, 45 (3), pp. 598-607.

VELASCO, I.; C. SPENCE y J. NAVARRA (2011): «El sistema perceptivo: esa pequeña máquina del tiempo», *Anales de Psicología*, 27 (1), pp. 195-201.

WALDEN, B. E.; R. A. PROSEK, A. A. MONTGOMERY, C. K. SCHERR y C. J. JONES (1977): «Effect of training on the visual recognition of consonants», *Journal of Speech and Hearing Research*, 20 (1), pp. 130-145.

WALKER, S.; V. BRUCE y C. O'MALLEY (1995): «Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect», *Perception & Psychophysics*, 57 (8), pp. 1124-1133.

YONOVITZ, A.; J. T. LOZAR, C. THOMPSON, D. R. FERRELL y M. ROSS (1977): «Fox-box illusion»: Simultaneous presentation of conflicting auditory and visual CV's», *The Journal of the Acoustical Society of America*, *62* (S1), S3.