Report

Current Biology

A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements

Highlights

- Visual cortex tracks better forward than backward unheard acoustic speech envelope
- Effects not trivially caused by correlation of visual with acoustic signal
- Stronger top-down control of the visual cortex during forward display of lip movements
- Top-down influence correlates with visual cortical entrainment effect

Authors

Anne Hauswald, Chrysa Lithari, Olivier Collignon, Elisa Leonardelli, Nathan Weisz

Correspondence

anne.hauswald@sbg.ac.at (A.H.), nathan.weisz@sbg.ac.at (N.W.)

In Brief

Successful lip-reading requires a visual-phonological transformation. Hauswald et al. show that, during the processing of silently played lip movements, the visual cortex tracks the missing acoustic speech information when played forward as compared to backward. The effect is under top-down control.

Hauswald et al., 2018, Current Biology *28*, 1453–1459
May 7, 2018
https://doi.org/10.1016/j.cub.2018.03.044

CellPress

Current Biology
Report

A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements

Anne Hauswald,[1,2,4,5,*] Chrysa Lithari,[1,2,4] Olivier Collignon,[2,3] Elisa Leonardelli,[2] and Nathan Weisz[1,2,*]
[1]Centre for Cognitive Neurosciences, University of Salzburg, Salzburg 5020, Austria
[2]CIMeC, Center for Mind/Brain Sciences, Università degli studi di Trento, Trento 38123, Italy
[3]Institute of Research in Psychology & Institute of NeuroScience, Université catholique de Louvain, Louvain 1348, Belgium
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: anne.hauswald@sbg.ac.at (A.H.), nathan.weisz@sbg.ac.at (N.W.)
https://doi.org/10.1016/j.cub.2018.03.044

SUMMARY

Successful lip-reading requires a mapping from visual to phonological information [1]. Recently, visual and motor cortices have been implicated in tracking lip movements (e.g., [2]). It remains unclear, however, whether visuo-phonological mapping occurs already at the level of the visual cortex–that is, whether this structure tracks the acoustic signal in a functionally relevant manner. To elucidate this, we investigated how the cortex tracks (i.e., entrains to) absent acoustic speech signals carried by silent lip movements. Crucially, we contrasted the entrainment to unheard forward (intelligible) and backward (unintelligible) acoustic speech. We observed that the visual cortex exhibited stronger entrainment to the unheard forward acoustic speech envelope compared to the unheard backward acoustic speech envelope. Supporting the notion of a visuo-phonological mapping process, this forward-backward difference of occipital entrainment was not present for actually observed lip movements. Importantly, the respective occipital region received more top-down input, especially from left premotor, primary motor, and somatosensory regions and, to a lesser extent, also from posterior temporal cortex. Strikingly, across participants, the extent of top-down modulation of the visual cortex stemming from these regions partially correlated with the strength of entrainment to absent acoustic forward speech envelope, but not to present forward lip movements. Our findings demonstrate that a distributed cortical network, including key dorsal stream auditory regions [3–5], influences how the visual cortex shows sensitivity to the intelligibility of speech while tracking silent lip movements.

RESULTS

Successful lip-reading in the absence of acoustic information requires mechanisms of mapping from the visual information to the corresponding but absent phonological code [1]. We know that the visual and motor cortices track lip movements for congruent compared to incongruent audiovisual speech [2], but the large-scale neural processes precisely involved in linking visual speech (lip movements) processing with the auditory content of the speech signal have remained obscure. We performed a magnetoencephalography (MEG) experiment in which 24 participants were exposed to silent lip movements corresponding to forward and backward speech. In a parallel behavioral experiment with 19 participants, we demonstrate that silent forward lip movements are intelligible while backward presentation is not: participants could correctly identify words above chance level when presented with silent forward visual speech, while performance for silent backward visual speech did not differ from chance level. We compared the neural tracking of the unheard acoustic speech by contrasting the coherence between (unheard) forward and backward acoustic speech envelopes with the brain activity elicited by silently presented forward and backward lip movements. Uncovering visual cortical regions via this analysis, we then performed Granger causality analysis to identify the cortical regions mediating top-down control, and we assessed to what extent this was correlated to the aforementioned entrainment effect. Importantly, we also analyzed occipital forward-backward entrainment and Granger causality for coherence between lip contour and brain activity during visual speech to show that the findings are specific for the envelope of (unheard) acoustic speech-brain coherence.

Audiovisual Coherence Peaks at 5 Hz for Forward and Backward Presentations

A short example of the lip signal and the corresponding audio signal as well as its envelope is depicted in Figure 1A. The coherence spectrum between lip contour and acoustic speech, lip-speech coherence (Figure 1B), exhibits a distinct peak at 5 Hz, matching well the syllable rhythm of speech [7, 8]. Contrasting forward and backward lip-speech coherence for the whole frequency band (1–12.5 Hz) did not reveal any differences (t test, all p values > 0.05). Comparing the forward-backward power spectra of lip contour and acoustic speech (AS) envelope separately for the relevant range (4–7 Hz) did not reveal any differences (t test, p values > 0.5).

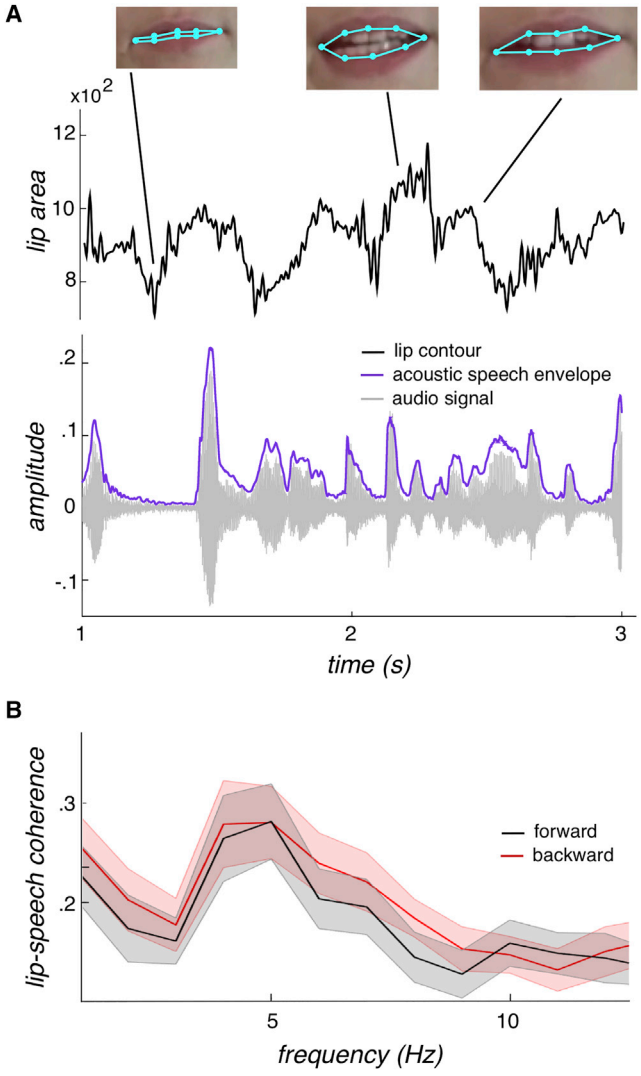

Figure 1. Lip Contour and Acoustic Speech Envelope

(A) The time series of the lip contour (area defined by the light blue points measured in square pixels) with the corresponding video frames is displayed together with the audio signal and the acoustic speech envelope [6] for a 2-s forward section. For the coherence analyses, the lip contour and the acoustic speech envelope were used.

(B) The coherence spectrum (averaged across stimuli) between lip contour and acoustic speech envelope is plotted peaking at 5 Hz, the rhythm of the syllables, for both forward and backward speech. The shaded area reflects the SD across stimuli. No difference between forward and backward speech occurred. For stimulus examples, see Videos S1 and S2.

Forward Presentations of Silent Lip Movements Are More Intelligible Than Backward Presentations

To ensure that silent lip movements differ in terms of intelligibility when presented forward or backward, we performed a behavioral experiment with separate participants than in the MEG experiment. Participants watched short videos of silent visual speech and were then asked to choose between two words, one of which was contained in the short video. Hits (mean: 64.84%) in the forward condition were significantly higher than chance level ($t(18) = 7.81$, $p < 0.0005$), while this was not the case for hits in the backward condition (mean: 53.47%, $t(18) = 1.54$, $p = 0.14$). Hits in the forward condition were also significantly higher than hits in the backward condition ($t(18) = 3.76$, $p < 0.005$; Figure 2A).

Stronger Entrainment between Occipital Activity and the Envelope of Unheard Acoustic Speech during Forward Presentation of Lip Movements

We first calculated the coherence between the absent acoustic envelope of the speech signal with the source-reconstructed (linearly constrained minimum variance, *LCMV*) MEG data on each voxel while participants were watching the silent lip movements. Occipital regions showed a statistically significant difference between the neural response to forward versus backward unheard acoustic speech envelope ($t(23) = 6.83$, $p < 0.000005$; Figures 2B and 2C). Given the high coherence between the acoustic and the visual signals related to speech stimuli (see above), the aforementioned effect could be a trivial by-product of a differential entrainment to the visual information only.

To test this possibility, we investigated the coherence between occipital activity and the lip contour (lip-brain coherence). The forward-backward difference was bigger ($t(23) = 3.43$, $p < 0.005$) for the unheard acoustic speech-brain coherence carried by lip movements than for lip-brain coherence elicited by lip movements ($t(23) = -0.07$, $p = 0.94$; Figure 2C; see Figure S1A for whole-brain contrast). Further, we contrasted the occipital coherence values of each condition with its respective surrogate data in which the time axis of the external signal (lip contour or unheard acoustic speech envelope) was flipped. This revealed that forward unheard acoustic speech-brain coherence as well as forward and backward lip-brain coherence, but not the backward unheard acoustic speech-brain coherence, were increased compared to their corresponding surrogate data (Figure S2). The same pattern was shown by the grand averages of all four conditions (Figure S3). This implies that the visual cortex tracks lip movements faithfully regardless of whether they are displayed forward or backward. However, only forward presented lip movements additionally elicit an entrainment of visual cortical activity to the envelope of the corresponding (unplayed) acoustic signal.

Top-Down Modulation on Visual Cortex Drives Unheard Acoustic Speech Envelope Entrainment Effects

To elucidate the network driving this putative visuo-phonological mapping process, we calculated Granger causality between the occipital region showing the strongest difference between forward and backward acoustic speech entrainment and the remaining whole-brain voxels. We focused the statistical analyses on relevant regions (parietal, temporal, and pre- and postcentral areas in the left hemisphere as defined by the Automatic Anatomical Labeling [AAL] atlas [9]) that broadly cover regions of interest as motivated by dual-route models of speech processing [3–5]. The contrast (corrected for multiple comparisons; see the STAR Methods) between Granger causality for forward and backward visual speech revealed increased Granger causality for the forward condition in left premotor, primary motor, and primary somatosensory cortex (Figure 3A). At a descriptive level, also posterior portions of the left superior temporal cortex (BA 22), inferior temporal gyrus (BA 37), and middle temporal

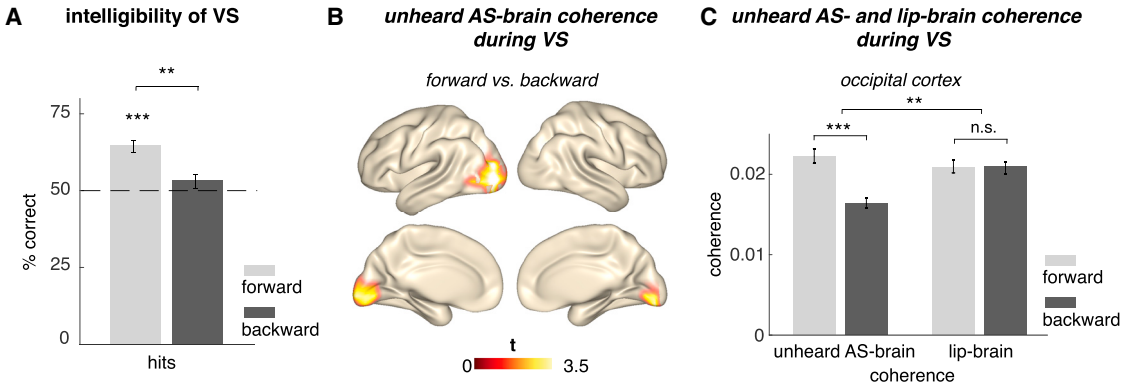

Figure 2. Behavioral and Entrainment Effects

(A) Separate behavioral experiment on intelligibility in silent visual speech (VS) with 19 participants. The contrast hits versus chance level was significant only in the forward condition. Hits in the forward condition were also higher than hits in the backward condition.

(B) Coherences of theta-band brain sources (4–7 Hz) with the not-heard acoustic envelope of speech (AS, while watching visual speech; VS, contrasted between forward and backward conditions; $p < 0.05$, Monte Carlo corrected) were increased at occipital regions when watching lip movements of forward speech compared to lip movements of backward speech.

(C) Mean of the individual unheard acoustic speech-brain and lip-brain coherence values during forward and backward presentations of visual speech extracted at the voxels of the statistical effect found in (B). Difference in occipital cortex between forward and backward unheard acoustic speech-brain coherence ($p < 0.000005$) was statistically bigger ($p < 0.005$) than the forward-backward difference of lip-brain coherence during visual speech (n.s., not significant). Error bars indicate SE. **$p < 0.005$ and ***$p < 0.0005$. For supporting analyses, see also Figures S1–S3.

gyrus (BA 39 including the angular gyrus) were above the uncorrected statistical critical value (see the STAR Methods). Altogether, this means that key nodes of mainly dorsal route processing regions exert relatively more top-down influence on the visual cortex during forward compared to backward presentation of visual speech.

To clarify whether the network-level effect was actually related to the unheard acoustic speech-brain coherence or a mere by-product of differential lip-brain coherence, we calculated the correlations of the forward Granger causality with both forward acoustic speech-brain coherence and lip-brain coherence in occipital regions. Only the correlations with unheard acoustic speech-brain coherence revealed significant results, while the lip-brain coherence did not ($p > 0.4$). For the unheard acoustic speech-brain coherence, mainly precentral and postcentral regions revealed a strong correlation (corrected for multiple comparisons, $p < 0.05$). At an uncorrected level, the premotor (BA 6), frontal eye field (BA 8), and posterior middle temporal regions (BA 39) also yielded correlations above the statistical critical value (see Figure 3B, left). The scatterplot in Figure 3B (right) illustrates this correlation for a precentral region ($r = 0.46$, $p < 0.05$) that showed a statistical effect for the forward-backward Granger causality contrast during visual speech (Figure 3A) and for the correlation between unheard acoustic speech-brain coherence and Granger causality (Figure 3B, left). Interestingly, these findings show a high correspondence with the relevant regions of the dual-route model of speech [3] (see Figure 3C).

DISCUSSION

Dual-process models of speech processing state the presence of two different routes [3–5]. While the ventral stream is assumed to contribute to comprehension, the dorsal stream is proposed to "map acoustic speech signals to frontal lobe articulation" [3]. Extending this notion, Rauschecker [4] proposes the dorsal stream to be a "supramodal reference frame," enabling flexible transformation of sensory signals (see also [10]). Most evidence underlying the development of these models used acoustic signals (e.g., [11, 12]).

Already very early in life, an audiovisual link between observing lip movements and hearing speech sounds is present, which consequentially supports infants to acquire their first language. In this context, visual speech constitutes a crucial role for speech processing, as can be seen by findings of infants of just 4–6 months of age who can discriminate their native language from silent lip movements only [13] or who allocate more attention to the mouth compared to the eyes once they detect synchrony between lip movements and speech sounds [14]. The importance of lip-reading for speech processing is also demonstrated by studies with deaf individuals [15], showing that lip-reading alone can be sufficient for language comprehension and suggesting a functional role that goes beyond the mere support of auditory input processing. In a recent study, Lazard and Giraud [1] hypothesized that the functional role of lip-reading was to enable the visual cortex to remap visual information into phonological information. Given the described mapping processes, we ask the following question: how does the mapping of an acoustic signal influence lip-reading when the acoustic signal is absent and only induced by the lip movements. We investigated this via cortical entrainment to the absent acoustic speech signal that was carried by the silent lip movements. We contrasted forward and backward visual conditions of speech segments, given that only the forward condition was intelligible, as shown by the behavioral experiment, and therefore should induce speech-related processes.

The auditory cortex has been repeatedly found to be activated and entrained in response to audio-only [16] or audiovisual stimulation [17, 18]. Previous fMRI studies established also

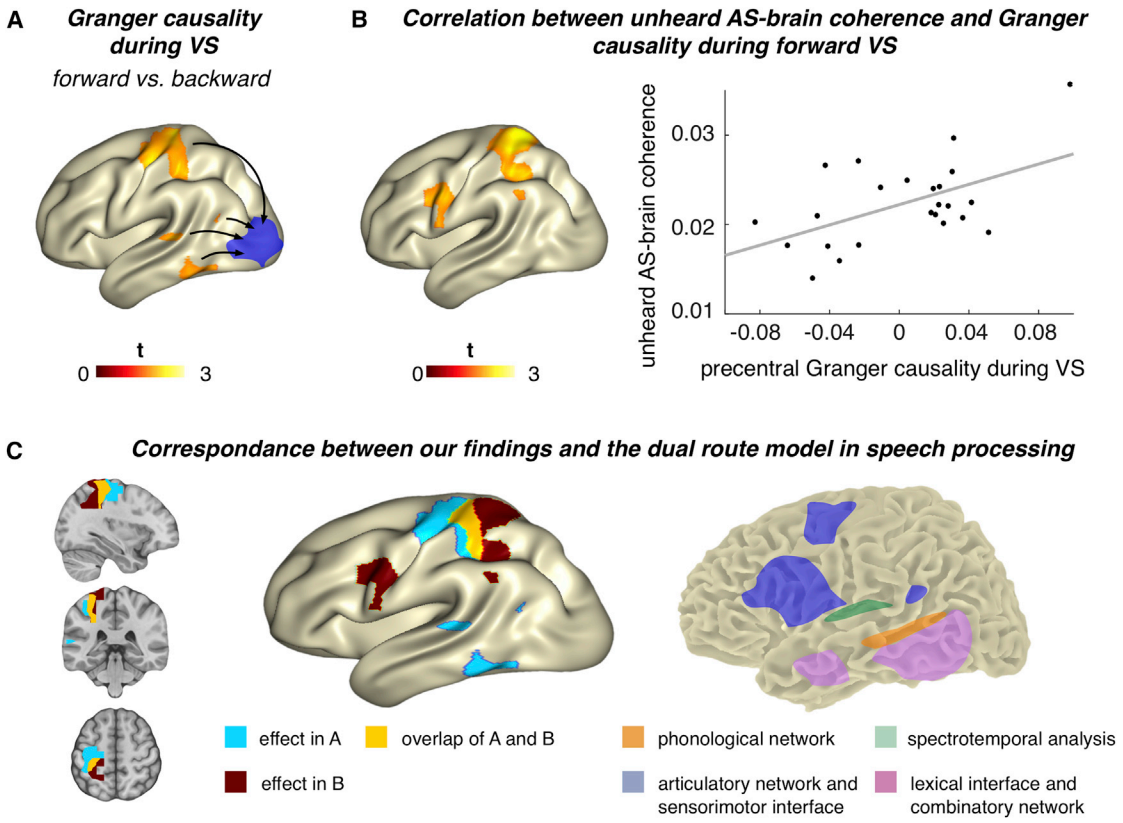

Figure 3. Top-Down Modulation of Visual Cortex during Visual Speech

(A) Granger causality during forward versus backward lip movements (visual speech, VS) using occipital seed region taken from an effect found for the envelope of unheard acoustic speech (AS)-brain coherence (blue) that was defined as voxel with the strongest forward-backward entrainment effect to unheard acoustic speech envelope. Calculation of contrast between normalized ([ingoing − outgoing]/[ingoing + outgoing]) forward and backward visual speech shows positive effects in the left premotor, primary motor, and primary somatosensory cortex (corrected for multiple comparisons) and (uncorrected) posterior superior (BA 22) and inferior temporal gyrus (BA 37) as well as the posterior middle temporal gyrus, including the angular gyrus (BA 39). Maps are masked by uncorrected statistical critical value (1.714).

(B) Significant correlation between the occipital unheard acoustic speech-brain coherence and the Granger causality of forward visual speech. Left: effects corrected for multiple comparison in pre-/postcentral gyrus. Also, above the uncorrected statistical critical value were values in inferior premotor cortex (BA 6), frontal eye fields (BA 8), and posterior middle temporal gyrus (BA 39). Right: scatterplot of correlation ($r = 0.46$, $p < 0.05$) in pre-/postcentral gyrus is shown.

(C) Illustration of the correspondence between our findings and the proposed dual-route model of speech processing [3]. Left: illustration shows the effects above the uncorrected statistical critical value found for differences in Granger causality between forward and backward visual speech (blue, see also A), correlation between unheard acoustic speech-brain coherence and Granger causality of forward visual speech (B, red), and their overlap (yellow) in anatomical and surface view. Right: illustration shows the dual-route model as proposed by Hickock and Poeppel (adapted from [5]) comprising the dorsal route (blue) and the ventral route (red).

that perception of silent visual speech activates the auditory cortices, clearly showing involvement of auditory processes during lip-reading [19, 20]. Recently, a study attempted a reconstruction of posterior surface electroencephalogram (EEG) channels (presumably capturing visual cortex activity) via the absent acoustic speech envelope [21]. The authors showed that a model based on the absent acoustic signal predicted posterior EEG activity similar to models based on frame-to-frame motion changes and categorical visual speech features [21]. However, since the acoustic signal was seen as a proxy for lip movements (which were not explicitly investigated), the separate contributions of acoustic and visual information were not explored. Going beyond this finding, we investigated cortical entrainment to the absent acoustic signal by comparing forward and backward envelopes of unheard acoustic speech carried by silent presentations of lip movements putatively linked to altered speech intelligibility (see Figure 2A). Using source-level analysis, we found that the visual cortex showed stronger entrainment (higher coherence) to unheard acoustic speech envelope during forward (intelligible) rather than backward (unintelligible) mute lip movements. Importantly, our control analysis of coherence between brain activity and the lip contour of the actually observed lip movements did not reveal similar forward versus backward differences (Figure 2C). This excludes the possibility that our findings for unheard acoustic speech-brain coherence are just an epiphenomenon, given that lip and audio signals are highly correlated (Figure 1B; *cf.* [22]).

Although the absence of an effect in lip-brain coherence might be initially surprising, it presumably is due to the lack of a specific task and suggests that, for a difference in forward-backward

lip-brain coherence in visual speech, targeted attention is needed while the putative visuo-phonological transformation occurs relatively automatically. For example, Park et al. [2] showed that the coherence between lip contour and left visual regions during audiovisual speech is modulated by attention. Further, this control analysis argues against the possibility that spontaneous attentional processes produced the difference in forward-backward unheard acoustic speech-brain coherence, as for both measures (unheard acoustic speech-brain coherence and lip-brain coherence) identical datasets of MEG recordings were used. An alternative interpretation of our finding could be based on a much better representation of the syllable structure in the acoustic signal compared to the visual signal. However, if our findings were simply due to a higher richness of the syllabic structure in the acoustic signal, this should also be reflected in the brain coherence with the backward unheard AS condition. In this condition, even though the onset dynamics change, the representation of the syllabic structure (expressed, e.g., by theta power) is the same. To our knowledge, this is the first time that an effect of speech intelligibility on brain coherence with unheard acoustic speech has been reported.

It was also recently shown that the visual cortex is important for tracking speech-related information, for example, sign language [23] and lip movements [2]. Further, the more adverse the condition (low signal-to-noise ratio), the more the visual cortex is entrained to the speech signal of actual acoustic speech presented together with varying levels of acoustic noise and either informative or uninformative visual lip movements at low frequencies [22]. Our study confirms and extends these findings by investigating how occipital regions track the acoustic envelope related to silent visual speech delivered by lip movements. We show that, in the absence of acoustic information, the unheard auditory signal, not the lip movement, entrains the visual cortex differentially for intelligible and unintelligible speech. In this case, a visuo-phonological mapping mechanism needs to be in place, and our results showing entrainment of visual areas to (non-presented) acoustic speech may be a reflection of such a process. Given that, in the absence of an actual task, no difference in forward/backward unheard acoustic speech envelope in auditory regions occurred, we propose that, while the putative visuo-phonological mapping process is automatic, it does not imply that the transformed information is necessarily used by auditory regions in a task-irrelevant context. Rather, this mapping presumably interacts with top-down processes as lip-reading has been reported to benefit from contextual cues [24].

Recent studies provide evidence for top-down processes in audiovisual and audio-only settings. For example, Giordano and colleagues [22] showed an increase in directed connectivity between superior frontal regions and visual cortex under the most challenging (acoustic noise and uninformative visual cues) conditions. Kayser et al. [25] proposed top-down processes modulating acoustic entrainment. Park et al. [26] showed enhanced top-down coupling between frontal and, in their case, auditory (due to only auditory stimuli) regions during intelligible speech in the left hemisphere compared to unintelligible speech. Going one step further, given the complete absence of auditory input during silent visual speech, we also expected similar enhanced top-down control to differentiate intelligibility but, in our case, of the visual cortex. Indeed, calculating Granger causality (4–7 Hz, forward and backward visual speech) between visual regions and the other regions of interest showed differential ingoing and outgoing connections for the two conditions. We contrasted the two normalized ([ingoing − outgoing]/[ingoing + outgoing]) conditions, and, as expected, the forward condition yielded a more positive ratio of ingoing and outgoing connections than the backward condition, stemming from mainly left (pre)motor regions and primary sensory regions. Also in auditory or audiovisual studies [2, 22, 25, 26], motor regions play an important role in top-down control.

Further, the posterior portions of left superior, middle, and inferior temporal gyrus show differences in the statistical comparison, although not significant at a cluster-corrected level. They are, nevertheless, reported because they provide interesting insights given their previously established role in speech processing [11], particularly under adverse conditions [27] and for audiovisual integration, respectively (overview in [28, 29]).

Importantly, the significance of the top-down processes for visuo-phonological mapping is further supported by the correlations between the acoustic speech-brain coherence in occipital regions during silent visual speech and the Granger causality effect. We find positive correlations mainly for precentral and postcentral regions but also statistically uncorrected at premotor areas (BA 6), the frontal eye field (BA 8), and the posterior middle temporal gyrus (BA 39). The precentral and postcentral areas showed a strong overlap with the regions that had enhanced Granger causality in the forward condition. The positive correlations in these regions suggest that the extent of top-down influence on visual cortical regions is associated, at least partially, with the magnitude of this region to exhibit entrainment to (not presented) acoustic speech input. Importantly, the Granger causality effects are not a by-product of the entrainment to the lip movements, as shown by the missing correlations between occipital lip-brain coherence and Granger causality in the regions of interest (relevant for the dual-route model). Again, this lack of effect might be due to the passive nature of our study, suggesting that lip-brain coherence (at least in visual areas) does not automatically couple with top-down auditory regions in the absence of an active task.

Conclusions

Our findings suggest that, while observing lip movements, acoustic speech synchronizes with the visual cortex even if the auditory part of the speech input is physically not present. Importantly, this cortical mechanism is sensitive to intelligibility, while the same is not the case when looking at entrainment to the actual visual signal. Thus, while observing forward (and more intelligible) lip movements, the visual cortex additionally tracks the absent acoustic envelope. Our results strongly suggest dorsal stream regions, including motor-related areas, may mediate this visuo-phonological mapping process by exerting top-down control of the visual cortex.

Referring again to the initially mentioned dual-route model of speech processing [3–5], our results show a strikingly high correspondence of involved regions. This underlines the importance of these regions in processing speech-relevant information across modalities. Overall, our study supports the idea that in particular dorsal processing routes are activated by the

observation of silent lip movements, enabling a top-down controlled mapping of the visual signal into the absent acoustic signals [1]. This mapping might be achieved via functional dependencies between auditory and visual sensory systems that exist even in the earliest stages of sensory processing in humans and animals [30–32].

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Stimuli and experimental procedure
  - MEG recording
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Extraction of acoustic speech envelope and lip contour signal
  - MEG preprocessing
  - Coherence calculation
  - Granger causality
  - Statistical analysis
  - Behavioral experiment
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, one data file, and two videos and can be found with this article online at https://doi.org/10.1016/j.cub.2018.03.044.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Markus Van Ackeren for his help in extracting the speech envelope. The contributions of C.L. and N.W. were financed by the European Research Council (WIN2CON, ERC, StG 283404). O.C. is supported by the European Research Council (MADVIS, ERC, StG 337573).

AUTHOR CONTRIBUTIONS

N.W. and O.C. designed the experiment. E.L. ran the behavioral experiment. A.H. and C.L. analyzed the data. All authors wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 8, 2018
Revised: February 25, 2018
Accepted: March 20, 2018
Published: April 19, 2018

REFERENCES

1. Lazard, D.S., and Giraud, A.L. (2017). Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome. Nat. Commun. *8*, 14872.
2. Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. eLife *5*, e14521.
3. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.
4. Rauschecker, J.P. (2012). Ventral and dorsal streams in the evolution of speech and language. Front. Evol. Neurosci. *4*, 7.
5. Rauschecker, J.P. (1998). Cortical processing of complex sounds. Curr. Opin. Neurobiol. *8*, 516–521.
6. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. PLoS Comput. Biol. *5*, e1000436.
7. Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. *3*, 238.
8. Giraud, A.L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. *15*, 511–517.
9. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage *15*, 273–289.
10. Rauschecker, J.P. (2018). Where, When, and How: Are they all sensorimotor? Towards a unified view of the dorsal pathway in vision and audition. Cortex *98*, 262–268.
11. Buchsbaum, B.R., Hickok, G., and Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cogn. Sci. *25*, 663–678.
12. Rauschecker, J.P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. Science *268*, 111–114.
13. Weikum, W.M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., and Werker, J.F. (2007). Visual language discrimination in infancy. Science *316*, 1159.
14. Lewkowicz, D.J., and Hansen-Tift, A.M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. Proc. Natl. Acad. Sci. USA *109*, 1431–1436.
15. Andersson, U., and Lidestam, B. (2005). Bottom-up driven speechreading in a speechreading expert: the case of AA (JK023). Ear Hear. *26*, 214–224.
16. Lalor, E.C., and Foxe, J.J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. *31*, 189–193.
17. Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. *8*, e1000445.
18. Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., and Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci. Lett. *127*, 141–145.
19. Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Tarkiainen, A., and Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport *16*, 125–128.
20. Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., and David, A.S. (1997). Activation of auditory cortex during silent lipreading. Science *276*, 593–596.
21. O'Sullivan, A.E., Crosse, M.J., Di Liberto, G.M., and Lalor, E.C. (2017). Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. Front. Hum. Neurosci. *10*, 679.
22. Giordano, B.L., Ince, R.A.A., Gross, J., Schyns, P.G., Panzeri, S., and Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. eLife *6*, e24763.
23. Brookshire, G., Lu, J., Nusbaum, H.C., Goldin-Meadow, S., and Casasanto, D. (2017). Visual cortex entrains to sign language. Proc. Natl. Acad. Sci. USA *114*, 6352–6357.

24. Spehar, B., Goebel, S., and Tye-Murray, N. (2015). Effects of Context Type on Lipreading and Listening Performance and Implications for Sentence Processing. J. Speech Lang. Hear. Res. *58*, 1093–1102.

25. Kayser, S.J., Ince, R.A.A., Gross, J., and Kayser, C. (2015). Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. J. Neurosci. *35*, 14691–14701.

26. Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., and Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol. *25*, 1649–1653.

27. Ozker, M., Schepers, I.M., Magnotti, J.F., Yoshor, D., and Beauchamp, M.S. (2017). A Double Dissociation between Anterior and Posterior Superior Temporal Gyrus for Processing Audiovisual Speech Demonstrated by Electrocorticography. J. Cogn. Neurosci. *29*, 1044–1060.

28. Bernstein, L.E., and Liebenthal, E. (2014). Neural pathways for visual speech perception. Front. Neurosci. *8*, 386.

29. Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 1001–1010.

30. Ghazanfar, A.A., and Schroeder, C.E. (2006). Is neocortex essentially multisensory? Trends Cogn. Sci. *10*, 278–285.

31. Murray, M.M., Lewkowicz, D.J., Amedi, A., and Wallace, M.T. (2016). Multisensory Processes: A Balancing Act across the Lifespan. Trends Neurosci. *39*, 567–579.

32. Lewkowicz, D.J. (1996). Perception of auditory-visual temporal synchrony in human infants. J. Exp. Psychol. Hum. Percept. Perform. *22*, 1094–1106.

33. Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. *2011*, 156869.

34. Smith, Z.M., Delgutte, B., and Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature *416*, 87–90.

35. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.

36. Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., and Anderson, C.H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. J. Am. Med. Inform. Assoc. *8*, 443–459.

37. Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. *11*, e1001752.

38. Tian, Y., Kanade, T., and Cohn, J. (2000). Robust lip tracking by combining shape, color and motion. In Proceedings of the 4th Asian Conference on Computer Vision, pp. 1040–1045.

39. Tomasi, C. (1991). Detection and tracking of point features. Carnegie Mellon University Technical Report, CMU-CS-91-132, Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.5770.

40. Van Veen, B.D., van Drongelen, W., Yuchtman, M., and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. *44*, 867–880.

41. Mattout, J., Henson, R.N., and Friston, K.J. (2007). Canonical source reconstruction for MEG. Comput. Intell. Neurosci. *2007*, 67613.

42. Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. Phys. Med. Biol. *48*, 3637–3652.

43. Dhamala, M., Rangarajan, G., and Ding, M. (2008). Analyzing information flow in brain networks with nonparametric Granger causality. Neuroimage *41*, 354–362.

44. Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron *85*, 390–401.

45. Michalareas, G., Vezoli, J., van Pelt, S., Schoffelen, J.M., Kennedy, H., and Fries, P. (2016). Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. Neuron *89*, 384–397.

46. Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB	The MathWorks	https://www.mathworks.com/products/matlab.html
Fieldtrip toolbox	[33]	http://www.fieldtriptoolbox.org/
Chimera toolbox	[34]	http://research.meei.harvard.edu/chimera/More.html
Psychtoolbox	[35]	http://psychtoolbox.org/
Caret	[36]	http://brainvis.wustl.edu/wiki/index.php/Caret:About

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Anne Hauswald (anne.hauswald@sbg.ac.at).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Twenty-four healthy volunteers (age 28.31 ± 4.6, 9 females, all right handed) with normal hearing and vision participated in the study. All participants were native Italian speakers. They gave their written informed consent and received 30 euros at the end of the experiment. Ethical approval was obtained from the University of Trento Ethics Committee and conducted according to the Declaration of Helsinki.

METHOD DETAILS

Stimuli and experimental procedure

The videos (visual speech) were recorded with a digital camera (VP-D15i; Samsung Electronics) at a rate of 25 frames per second. The audio files were recorded at a sampling rate of 44.1 kHz. The speakers were native Italians (two males and one female). Nine short text pieces were recorded from each speaker lasting from 21 to 40 s each, resulting in 27 forward videos and audio files and, by reversing them, in 27 backward video and audio files (see Data S1 for text examples as well as Videos S1 and S2). On average 36 pieces were randomly selected and presented to each participant counterbalancing the gender of the speakers. The mute videos were displayed on a projector panel in the MEG chamber and the audio files were presented binaurally via a sound pressure transducer (Etymotic Research ER-2) through two plastic tubes terminating in plastic earplugs while participants were fixating on a cross at the center of the projector panel. The order of the visual and the auditory sessions was counterbalanced. Participants were instructed to passively watch the mute videos and listen to the audible speech. The experiment lasted less than one hour including preparation. Presentation was controlled via Psychtoolbox [35].

MEG recording

MEG was recorded at a sampling rate of 1 kHz using a 306-channel (204 first order planar gradiometers) VectorView MEG system (Elekta-Neuromag, Helsinki, Finland) in a magnetically shielded room (AK3B, Vakuumschmelze, Hanau, Germany). MEG signal was online high-pass and low-pass filtered at 0.1 Hz and 330 Hz respectively. Prior to the experiment, individual head shapes were digitized for each participant including fiducials (nasion, pre-auricular points) and around 300 points on the scalp using a Polhemus Fastrak digitizer (Polhemus, Vermont, USA). The head position relative to the MEG sensors was continuously controlled within a block through five head position indicator (HPI) coils (at frequencies: 293, 307, 314, 321 and 328 Hz). Head movements did not exceed 1.5 cm within and between blocks.

QUANTIFICATION AND STATISTICAL ANALYSIS

Extraction of acoustic speech envelope and lip contour signal

The acoustic speech envelope was extracted using the Chimera toolbox by Delguette and colleagues (http://research.meei.harvard.edu/chimera/More.html) following a well-established approach in the field [6, 37] where nine frequency bands in the range of 100 – 10000 Hz were constructed as equidistant on the cochlear map [34]. Sound stimuli were band-pass filtered (forward and reverse to avoid edge artifacts) in these bands using a 4^{th}-order Butterworth filter. For each band, envelopes were calculated as

absolute values of the Hilbert transform and were averaged across bands to obtain the full-band envelope that was used for coherence analysis. The envelope was then down-sampled to 512 Hz to match the down-sampled MEG signal.

The lip contour of the visual speech was extracted with an in-house algorithm in MATLAB calculating the area (function *polyarea.m*) defined by eight points on the lips of the speaker (Figure 1A). These points were defined in the first frame of the video and were tracked using the Kanade-Lucas-Tomasi algorithm (KLT, function *vision.PointTracker*) on the next frames [38, 39]. The fluctuations of the mouth area are thus expressed in this signal at the sampling rate of the video (25 frames/sec), which was interpolated to 512 Hz to match the MEG signal.

MEG preprocessing

Data were analyzed offline using the Fieldtrip toolbox [33]. First, a high-pass filter at 1 Hz (6th order Butterworth IIR) was applied to continuous MEG data. Then, trials were defined keeping 2 s prior to the beginning of each stimulus and post-stimulus varying according to the duration of each stimulus. Trials containing physiological or acquisition artifacts were rejected. Bad channels were excluded from the whole dataset. Sensor space trials were projected into source space using linearly constrained minimum variance beamformer filters [40] and further analysis was performed on the obtained time-series of each brain voxel. The procedure is described in detail here: http://www.fieldtriptoolbox.org/tutorial/shared/virtual_sensors. To transfer the data into source space, we used a template structural magnetic resonance image (MRI) from Montreal Neurological Institute (MNI) and warped it to the subject's head shape (Polhemus points) to optimally match the individual fiducials and headshape landmarks. This procedure is part of the standard SPM (http://www.fil.ion.ucl.ac.uk/spm/) procedure of canonical brain localization [41].

A 3D grid covering the entire brain volume (resolution of 1 cm) was created based on the standard MNI template MRI. The MNI space equidistantly placed grid was then morphed to individual headspace. Finally, we used a mask to keep only the voxels corresponding to the gray matter (1457 voxels). Using a grid derived from the MNI template allowed us to average and compute statistics as each grid point in the warped grid belongs to the same brain region across participants, despite different head coordinates. The aligned brain volumes were further used to create single-shell head models and lead field matrices [42]. The average covariance matrix, the head model and the leadfield matrix were used to calculate beamformer filters. The filters were subsequently multiplied with the sensor space trials resulting in single trial time-series in source space. For both power and coherence, only the post-stimulus period was considered and the initial long stimulus-based trials were cut in trials of 2 s to increase the signal-to-noise ratio. The number of forward and backward trials was equalized (mean: 239.46 +- SD 14.97 for visual speech). The number of forward and backward trials did not differ statistically for the female and male speakers.

Coherence calculation

The cross-spectral density between the acoustic speech envelope and the corresponding lip contour was calculated on single trials with multitaper frequency transformation (dpss taper: 1-20 Hz in 1 Hz steps, 1 Hz smoothing). The same was then done between each virtual sensor and the acoustic speech envelope as well as the lip contour (dpss taper; 1 – 20 Hz in 1 Hz steps; 3 Hz smoothing). Then, the coherence between activity at each virtual sensor and the acoustic speech envelope or the lip contour while participants either watched the lip movements or heard the speech was obtained in the frequency spectrum and averaged across trials. We will refer to the coherence between acoustic speech envelope and brain activity as acoustic speech-brain coherence and between the lip-contour and brain activity as lip-brain coherence.

Granger causality

We took all voxels of the statistical effect (Figure 2A) between forward and backward unheard acoustic speech-brain coherence and from those identified for each participant an individual voxel based on the maximum difference in forward versus backward unheard acoustic speech-brain coherence that occurred. This voxel was then used as a seed for calculating Granger causality during video presentation (visual speech) with all other voxels [43]. Fourier coefficients were calculated using multitaper frequency transformation with a spectral smoothing of 3 Hz followed by calculating bivariate granger causality. This led to measures of ingoing and outgoing connections for the individual seed voxel for forward and backward visual speech which we normalized ((ingoing – outgoing)/(ingoing+ outgoing)) separately for forward and backward visual speech (see also [44, 45]). Note that in the following, whenever we speak of Granger causality we refer to normalized Granger causality with an individual occipital voxel as seed region.

Statistical analysis

To test for differences in source space, forward versus backward contrast was performed for acoustic speech-brain coherence on a spectrum level discarding the time dimension for the whole brain. A two-tailed dependent samples t test was carried out averaging the coherence values over our frequency band of interest, theta (4 – 7 Hz), as the lip-speech coherence showed a clear peak in this frequency. Consistent with Ghitza [7] and Giraud and Poeppel [8] in our stimulus material, this frequency corresponded with the frequency of syllables (mean = 5.05 Hz, range: 4.1-5.6 Hz). Note that when contrasting the forward and backward coherence between MEG activity and the acoustic speech envelope watching lip movements no acoustic signal was present. For the Granger causality, we also averaged values over our frequency band of interest, theta (4 – 7 Hz). Given findings of enhanced top-down coupling during intelligible speech compared to unintelligible speech in the left hemisphere [26], we expected more connectivity during the forward condition (intelligible) and therefore used a one-tailed t test as test statistic. Further, based on models of dual-route processing of speech [3, 4] proposing mapping of acoustic signals within specific regions of the left hemisphere, we took a more statistically

focused approach and accordingly selected all voxels within left temporal, parietal, postcentral and precentral regions as defined by the AAL atlas [9] implemented in Fieldtrip broadly covering the regions of interest as proposed by the dual-stream model [3].

To control for multiple comparisons, a non-parametric Monte-Carlo randomization test was undertaken [46]. The t test was repeated 5000 times on data shuffled across conditions and the largest t-value of a cluster coherent in space was kept in memory. The observed clusters were compared against the distribution obtained from the randomization procedure and were considered significant when their probability was below 5%. Significant clusters were identified in space. For contrasting forward and backward lip-speech coherence we used Monte.Carlo permutation with FDR for multiple comparisons correction.

For statistical comparison of the individually extracted values of occipital forward versus backward unheard acoustic speech-brain coherence and occipital lip-brain coherence, we used two-tailed t tests.

We correlated the effects from the unheard acoustic speech-brain coherence with the results of the Granger causality analysis. We calculated the mean across voxels from the occipital coherence effect during the forward visual speech and correlated this with the selected regions in left temporal, parietal, postcentral regions and precentral regions. While a significant effect was found, we also report the effects above the statistical critical value that did not survive cluster-correction, as the patterns overlap strongly with the regions obtained from the Granger contrast, adding support to their potential functional relevance. We looked at the conjunction between the effects from the forward-backward Granger causality contrast and the correlation analysis. Voxels in precentral and postcentral gyrus showed overlapping effects, and for the voxel with the strongest correlation we display the scatterplot illustrating the correlations between significant voxel of the forward-backward unheard acoustic speech-brain coherence contrast and the overlapping voxels of the Granger causality forward-backward contrast and the correlation analysis.

For visualization, source localizations of significant results were mapped onto inflated cortices using Caret [36] or a standard MNI brain as implemented in Fieldtrip.

Behavioral experiment

To elucidate if visual speech presented without sound also differs in terms of intelligibility, we performed an independent behavioral experiment with 19 Italian native speakers (age 32.4 ± 3.9, 12 females). We used the same stimuli as in the MEG experiment cut in phrases of 3.5 – 7.5 s duration (see Data S1 for text examples). The videos of the phrases (25 forward and 25 backward) were presented without sound and at the end of each trial, two words appeared on the screen and the participant responded by button press which of the two was contained in the short presented snippet. In both the forward and backward condition one option represented the correct word. Performance was statistically analyzed using t tests between hit rates of the forward and backward condition and between chance level (50%) and the hit rates.

DATA AND SOFTWARE AVAILABILITY

MEG raw data are available upon request by contacting the Lead Contact, Anne Hauswald (anne.hauswald@sbg.ac.at).

Current Biology, Volume *28*

Supplemental Information

A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements

Anne Hauswald, Chrysa Lithari, Olivier Collignon, Elisa Leonardelli, and Nathan Weisz

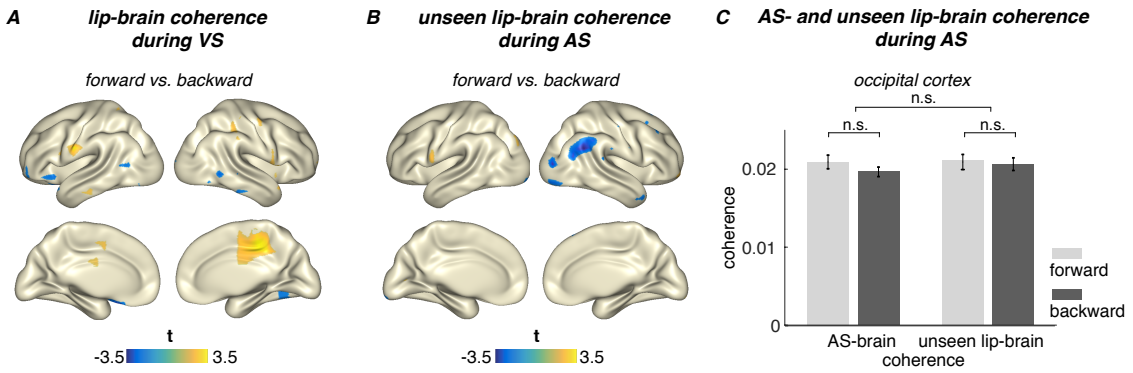

Figure S1: **Lip-brain coherence during VS and AS. Related to Figure 2B and C.** A) Coherences of theta band brain sources (4-7 Hz) with the lip contour of speech while watching visual speech (VS, contrasted between forward and backward conditions, $p < .05$, not corrected for multiple comparisons) is increased at paracentral regions, face regions of primary motor cortex, left inferior temporal regions, and decreased at bilateral inferior temporal regions and left frontal superior lobe. Importantly, no visual cortex effect between forward and backward lip-brain coherence during VS was identified, even at liberal statistical threshold. This underlines the uniqueness of our finding of increased coherence between unheard AS and activity in visual cortex. B) Coherences of theta band brain sources (4-7 Hz) with the not-seen lip contour while listening to acoustic speech (AS, contrasted between forward and backward conditions, $p < .05$, not corrected for multiple comparisons) is decreased during forward presentation mainly at right angular cortex. These findings support the uniqueness of our finding of visual regions showing increased coherence with unheard AS. C) Mean of the individual acoustic speech (AS)-brain and unseen lip-brain coherence values (extracted at the occipital voxels of the statistical effect found for forward versus backward unheard acoustic speech-coherence) during acoustic speech. Contrast of forward-backward acoustic speech-brain coherence ($t(23)=1.29$, $p=0.21$) and unseen lip-brain coherence ($t(23)=-0.43$, $p=0.67$) did not show differences. Also, the contrast between forward-backward difference for acoustic speech-brain and unseen lip-brain was not significant ($t(23)=0.76$, $p=0.43$). Error bars indicate standard error.

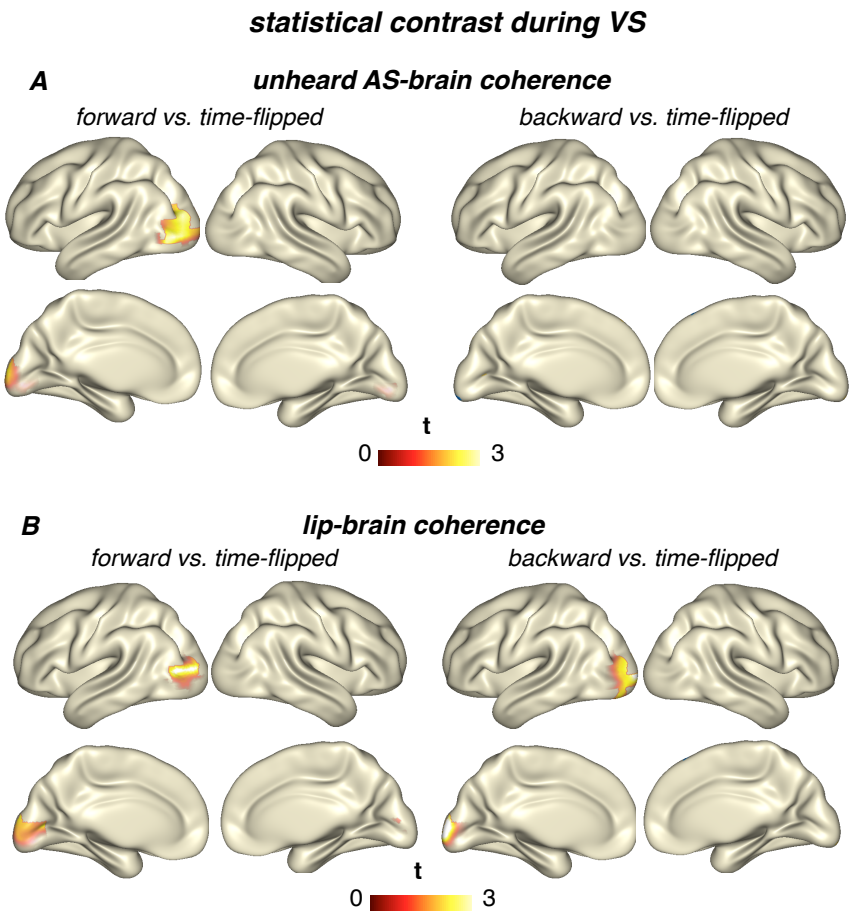

Figure S2: Unheard AS-brain and lip-brain coherence during VS contrasted with time-flipped data. Related to Figure 2B. Coherence between brain activity at 4-7 Hz and unheard acoustic speech envelope (AS, A) as well as lip contour (B) while watching visual speech contrasted with the respective time-flipped surrogate data. We tested if forward and backward coherences tracked the corresponding signal (lip for lip-brain coherence, acoustic speech envelope for unheard AS-brain coherence) compared to their surrogate data of time-flipped lip contour or unheard AS. Effects are calculated for the visual voxels showing the difference between forward and backward unheard AS-brain coherence (see. Figure 2B). As expected and correcting for multiple comparison (p<0.05, 1000 randomization) we found increased coherence compared to time-flipped surrogate data for forward lip-brain coherence (p=0.006), backward lip-brain coherence (p=0.01) and forward unheard AS-brain coherence (p=0.003) while no effect occurred for backward unheard AS-brain coherence compared to time-flipped surrogate data. This analysis supports our conclusion that visual cortex faithfully tracks both forward and backward lip movements during visual speech but additionally only tracks the unheard forward acoustic speech envelope.

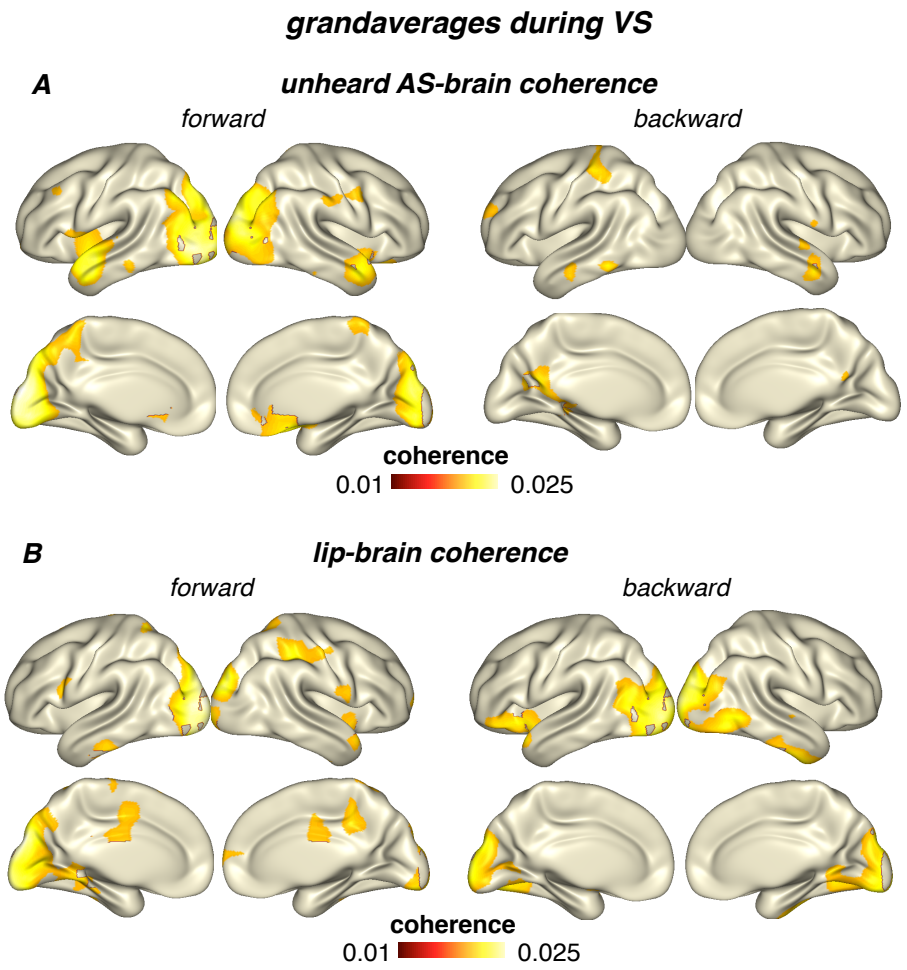

Figure S3: Grand averages of unheard As-brain and lip-brain coherence during VS. Related to Figure 2B and C. Grand averages of coherence between brain activity at 4-7 Hz and unheard acoustic speech envelope (AS, A) as well as lip contour (B) separately for forward and backward conditions while watching visual speech. Grand averages are masked by 75% of the maximum value of unheard AS-brain coherence (0.0198). Maximal coherence is consistently seen in occipital regions with the exception of unheard backward AS-brain coherence during VS. Furthermore, auditory regions also showed increased values during the forward unheard AS-brain coherence also regions, particularly, the superior temporal lobes showed increased coherence, including parts of the auditory cortex (Brodmann area 22) and Brodmann area 21. These results support our conclusion that the visual cortex tracks forward and backward lip movements as well as additionally the unheard AS during VS. Further, the grand averages show some involvement of auditory-related regions during the putative visuo-phonological transformation during the forward unheard AS condition.