Causal inference and temporal predictions in audiovisual perception of speech and music

Noppeney, Uta; Lee, Hwee Ling

DOI:
10.1111/nyas.13615

Causal inference and temporal predictions in audiovisual perception of speech and music

Uta Noppeney[1], Hwee Ling Lee[2]

[1]Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, UK

[2]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Corresponding author:
Dr. Hwee Ling Lee
German Center for Neurodegenerative Diseases (DZNE), Bonn
Sigmund-Freud-Strasse 27 53127 Bonn Germany
Tel: +49-228-43302-870
E-Mail: hwee-ling.lee@dzne.de

Short title: Audiovisual perception of speech and music

Keywords: Audiovisual; speech; music; prediction error; Bayesian causal inference

Abstract

To form a coherent percept of the environment, the brain should integrate sensory signals emanating from a common source, but segregate those from different sources. Temporal regularities are prominent cues for multisensory integration in particular for speech and music perception. In line with models of predictive coding, we suggest that the brain adapts an internal model to the statistical regularities in its environment. This internal model enables cross-sensory and sensorimotor temporal predictions as a mechanism to arbitrate between integration and segregation of signals from different senses.

Audiovisual integration and scene analysis as causal inference

Imagine you are at a concert of a large symphony orchestra. Your senses are overwhelmed by all the sounds coming from many different instruments: the soar of the violins, the blast of the trumpets, the whistle of the flutes, the deep resonance of the double basses and the pounding of the timpani. At one moment, you attend selectively to the melody played by the first violins and concurrently watch their bows moving smoothly across the strings. Suddenly, their melody is interrupted by the crash of timpani. Each time you see the mallet hitting the skin, you hear a big bang. How can you segregate and selectively attend to the part played by the violins or timpani that your eyes currently focus on? How does the brain bind this multitude of auditory and visual signals into a structured experience of a concert performance rather than a cacophonic multisensory chaos?

In order to transform the audiovisual signals into a unified percept, the brain needs to solve the so-called causal inference problem and determine whether auditory (e.g. the bang of the timpani) and visual signals (e.g. the musician's hand and arm movements that lead the mallets to hit the timpani) are caused by same or different instruments and musicians. Ideally, it should bind signals from vision and audition when they come from the same instrument but process them independently when they come from different sources. Prior knowledge, spatial, temporal and other higher order statistical correspondence cues can inform the brain whether or not auditory and visual signals come from common or independent sources [1-8]: First, the concert attendees can use prior knowledge to associate auditory signals with distinct instrumental groups based on their specific pitch and acoustic colours (e.g. the acoustic colour of violins and tympani are clearly discernible). Second, they can bind

signals from vision and audition based on them happening at the same place, i.e. spatial concordance [9-12]. If a sound played by a string instrument comes from the right-hand side, it is more likely to be produced by the celli that are located in the right wing of the orchestra than by the violins that sit in the opposite half. Third, even when pitch, acoustic colour and spatial location are not informative, the brain can infer whether auditory and visual signals emanate from common sources or events based on audiovisual synchrony or temporal signal correlations [8, 13, 14]. For instance, based on audiovisual temporal correlations we can selectively bind the sight of the bow movements of the first violins with the auditory melody they play and segregate it from counteracting tones played by the second violins. In summary, audiovisual scene analysis requires the brain to infer the causal structure that generates the sensory signals by combining top-down prior knowledge with a variety of temporal, spatial and higher order statistical congruency cues.

Bayesian framework to model audiovisual scene analysis

From a Bayesian perspective, the brain is thought to perform audiovisual scene analysis by forming a probabilistic generative model of the sensory inputs that is inverted during perceptual inference [15]. Bayesian probability theory provides a normative framework that formulates how observers should combine uncertain sensory information to form a representation of the world (e.g. multisensory percept of an orchestra performance). Recent models of Bayesian Causal Inference account for human multisensory integration performance by explicitly modelling the potential causal structures that could have generated the observed sensory signals [3, 11, 16-18], i.e. whether auditory and visual signals emanate from one common or two independent sources. Under the assumption of a common signal source, the two uni-

sensory estimates of a physical property (e.g. stimulus location, onset time, duration, shape etc.) are combined weighted according to their relative sensory reliabilities (= inverse of variance) [19-24]. Under the hypothesis of two different sources, the auditory and visual signals are processed independently. On a particular instance, the brain cannot directly access the causal structure of the world, i.e. whether signals come from common or independent sources. Instead, it needs to infer the causal structure from the noisy sensory signals themselves such as them happening at the same time or space. To account for this inherent uncertainty about the world's causal structure, a final estimate of the physical property in question (e.g. stimulus duration, timing) is obtained by combining the estimates of the physical property under various causal structures using decisional strategies such as model averaging, model selection or probability matching [18].

Indeed, numerous studies have shown that human observers arbitrate between audiovisual integration and segregation for speech and music processing in line with the principles of Bayesian Causal Inference. This has been illustrated both in i. explicit causal inference tasks where participants explicitly determine whether auditory and visual signals come from a common source or judge audiovisual discrepancy (e.g. temporal asynchrony, spatial disparity), and ii. implicit causal inference tasks where the influence of causal inference is characterized implicitly by measuring its effect on multisensory integration and perceptual inference [25]. In other words, even when participants do not explicitly judge the causal structure, their causal decision determines whether and how they integrate signals into an audiovisual percept of a property in the environment (e.g. spatial location, onset timing). For speech recognition, causal inference has been characterized most

extensively in the McGurk illusion where observers integrate an auditory ‘ba’ and visual ‘ga’ phoneme into an audiovisual ‘da’ percept [26]. Critically, observer’s perception of a common source and the emergence of an integrated audiovisual ‘da’ percept decrease with increasing audiovisual temporal asynchrony [5, 6, 8, 27]. In music, the ‘pluck and bow’ illusion is a related yet perhaps less well studied illusion [28]. In the ‘pluck and bow’ illusion, observers are presented with a video showing an actor plucking or bowing a cello and a range of auditory signals that morph successively from a sound of a pluck into a sound of a bow stimulus. Likewise, observers were able to monitor the perceptual discrepancy of the auditory and visual signals in explicit causal inference tasks. Further, the influence of the visual signal on the observer’s auditory ‘pluck’ or ‘bow’ percept was influenced by the discrepancy between auditory and visual pluck or bow signals. Another recent study demonstrated that visual gestures influence the estimation of sound duration for percussive, but not sustained sounds. Critically, the impact of visual gestures were observed only if the gesture preceded the sound by up to 700ms but not if the gesture succeeded the sound, thus indicating that these multisensory effects relied on causal inference and a temporal integration window [29]. Collectively this research suggests that causal inference depends on a range of correspondence cues such as temporal asynchrony, spatial disparity or other higher order statistical correspondences (e.g. phoneme congruency) and is critical for audiovisual perceptual inference.

Parametric models of Bayesian causal inference predict human behaviour well in classical experimental settings where observers are presented with a limited number of sensory signals and need to arbitrate between a small number of causal

structures, i.e. whether auditory and visual signals come from one or two sources (Fig. 1). Yet, they are likely to face difficulties accommodating the numerous potential causal structures underlying the sensory richness and complexity that is characteristic of real-world situations such as an orchestra performance with many different players and instruments. This suggests that parametric models of Bayesian causal inference may define the normative and computational principles underlying multisensory integration and audiovisual scene analysis, yet the brain will need non-parametric or approximate inference mechanisms or even simple heuristics to solve causal inference problems facing the brain in our natural environment [30, 31]. Critically, irrespective of the exact computational algorithms, the brain may use multisensory causal inference that relies on a range of correspondence cues that indicate whether signals in different senses are attributable to common events in the environment. Given the importance of temporal information for music and speech processing, we shall next discuss predictive coding and internal sensorimotor forward models as two complementary mechanisms that may allow the brain to determine whether sensory signals come from a common source based on cross-sensory or sensorimotor temporal predictions.

Fig. 1 about here

Predictive coding and temporal predictions across the senses

The theory of predictive coding posits that the human brain optimizes an internal model of its environment by reducing the errors between its top-down predictions and the bottom-up sensory inputs across multiple levels of the cortical hierarchy [32-34]. Backward connections provide predictions from higher to subordinate cortical levels.

Conversely, forward connections furnish the prediction error that is computed at each cortical level as the difference between top-down predictions and bottom-up inputs. Research to date has focused predominantly on predictive coding as a mechanism for perceptual inference in uni-sensory (e.g. auditory or visual) domains and showed that observer's top-down predictions shape how we form a perceptual interpretation of the incoming noisy sensory signals [35-40]. Yet, everyday experience with the multisensory world will tune the internal model also to the statistics of natural audiovisual stimuli. In particular, lifelong exposure to audiovisual speech and music stimuli will shape the cortical hierarchical architecture to recapitulate their complex temporal structure evolving concurrently in vision and audition (for higher order cross-sensory predictions e.g. phoneme, gender, semantics see e.g. [9, 41-45]). This internal model enables the brain to predict not only the temporal evolution of the visual and/or auditory speech or music inputs but also their temporal relationship, thereby imposing temporal constraints on audiovisual integration. Audiovisual signals that match observers' expectations or predictions about the relative timing of the sensory signals should be bound into a unified percept, while sensory signals that violate expectations should be perceived as subjectively asynchronous and hence be less likely to be integrated into a unified percept.

In line with the principles of predictive coding, a large body of research has shown that audiovisual binding depends on bottom-up audiovisual stimulus statistics and observers' top-down prior expectations [2, 7]. While simple transient auditory and visual signals (e.g. beeps and flashes) do not need to be precisely synchronous, they need to co-occur within a narrow temporal window of integration of tens of milliseconds in order to be bound into a unified percept [46-49]. By contrast, trains of brief audiovisual

signals are bound even when the individual beeps and flashes are not temporally co-incident but evolve in a temporally correlated fashion. As a result, continuous signals are integrated based on a shared temporal structure (e.g. as quantified by non-zero lag temporal correlations) leading to a broader window of integration [14, 50].

Critically, the width and shape of the temporal integration window is moulded by observers' prior expectations that adapt to the audiovisual statistics of the environment at multiple timescales. At a fast timescale, the temporal integration window and point of subjective simultaneity (i.e. the relative audiovisual timing that maximizes perceived simultaneity) rapidly adapts to the level of asynchrony of the audiovisual signals (for review see [51]). For instance, when presented with auditory leading signals, observers recalibrate the perceived simultaneity of the audiovisual signals such that auditory leading signals are more likely to be perceived as synchronous [52, 53]. At longer timescales, lifelong exposure to environmental sensory statistics shapes the temporal integration window for natural stimuli such as speech and music. Thus, the broad and asymmetric temporal integration window for speech stimuli has been attributed to the statistical regularities of audiovisual speech where the onset of the voice - at least at the beginning of an utterance - lags the timing of the mouth movements approximately between 100 and 300 ms [54]. To accommodate this audiovisual lag in natural speech, observers are less likely to perceive auditory leading stimuli synchronous than auditory lagging stimuli resulting in an asymmetric temporal binding window [8, 55, 56]. Indeed, when the auditory signal component is spectrally rotated such that the auditory envelope is mostly preserved, yet the speech stimulus is rendered unintelligible and novel, the temporal integration window is wider and less asymmetric [57]. Likewise, observers are faster to detect audiovisual

mismatches for utterances presented in their native than in their foreign language [58]. Furthermore, the audiovisual temporal binding window narrows with perceptual training [59-61]. Together these studies highlight that the brain flexibly attunes an internal model to the statistical regularities of the audiovisual inputs. This internal model enables more precise temporal predictions for speech and music that match the statistical regularities in their natural environment leading to a narrower temporal binding window and audiovisual benefits for naturalistic relative to transformed stimuli.

The internal model also enables the brain to make cross-sensory predictions operating from vision to audition and vice versa. In particular, as visible movements often precede the auditory signal in speech (e.g. facial movements) [8, 62] and music (e.g. arm movement of the drummer) actions [63-65], the brain can use the visual signal to predict the temporal evolution of the auditory signal. Sensory signals that are incongruent or are physically delayed to one another should therefore elicit a prediction error signal. Generally, it is thought that prediction errors are associated with an increase in neural activity in the gamma band that carry feed-forward influences and an enhanced blood oxygenated level dependent (BOLD) response [32, 66], though we note that task context and other higher cognitive factors can alter whether audiovisual incongruencies are associated with increases or decreases in BOLD response. In line with this conjecture, a human magnetoencelography (MEG) study demonstrated that auditory speech signals that are incongruent to the facial movements increase gamma oscillations in lower auditory regions indexing a bottom-up prediction error [67]. Moreover, the increase in gamma oscillations for non-matching auditory signals depended on the predictiveness of the facial movement.

As expected, it was strongest when the auditory signal was presented with a video that strongly predicted a different phoneme (for review see: [68]).

Along similar lines, a recent neuroimaging study demonstrated that temporal misalignment of auditory and visual signal components of speech and music stimuli induced activation increases signalling a prediction error in low-level audiovisual areas and the superior temporal sulcus as a key audiovisual integration region [69]. Critically, the regional expression of the temporal prediction error depended on the directionality of the temporal misalignment. For both speech and music stimuli, auditory leading asynchronous stimuli induced a prediction error signal predominantly in human visual motion area MT, while visual leading asynchronous stimuli induced a prediction error signal in auditory areas (Fig. 2). These results suggest that the sensory system of the leading signal generates temporal predictions that are violated by the lagging sensory signal, so that prediction error signals are generated predominantly in the sensory system dedicated to the processing of the lagging signal.

In summary, predictive coding may form a generic mechanism that enables the brain to predict the temporal structure and relative timing of inputs from multiple senses. These temporal predictions are critical for inferring whether or not sensory signals come from a common source and should be bound into a coherent percept of our environment.

Fig. 2 about here

Internal forward models for sensori-motor temporal predictions

So far we have focused on how sensory systems encode an internal model that progressively adapts to the temporal structure and correlations of sensory inputs. Critically, natural speech and music are generated by actions often performed by human agents. Further, it is well established that even passive speech and music perception implicitly activate parts of the action system [70-72]. Given these intimate perception-action links, the brain may be able to provide more precise temporal predictions for music and speech stimuli by harnessing the computational operations involved in motor timing [73].

In the field of motor control, precise motor timing is thought to rely on the formation of internal forward models that map from the motor plan of the intended action (e.g. singing, speaking or violin playing) onto its sensory consequences (e.g. the visible finger movements and concurrent auditory sounds) [74]. They are fine-tuned to specific motor tasks and effectors via error feedback during interactions with the environment and thought to be instantiated in a cortico-cerebellar circuitry [75-78]. As many actions such as speech and music produce 'sensory consequences' concurrently in multiple sensory modalities, this internal forward model indirectly also furnishes predictions about the relative timings of the sensory signals such as the sound and the visible hand or mouth movement. Critically, the precision of these temporal predictions during perception should depend on observer's motor expertise for the specific observed action such as piano playing. Only observers that are trained on the relevant motor repertoire should be able to generate more precise temporal predictions leading to a greater sensitivity to audiovisual temporal misalignments and a narrower temporal binding window. While most human observers are speech

experts, humans vary considerably in their musical expertise. This makes the musician's brain an ideal model to study the relationship between observer's cross-sensory temporal predictions, audiovisual perception and their motor (or music) expertise [72, 79, 80] (see Table 1i.a for a summary of studies that examined the effect of long-term music training on perception of audiovisual speech and/or music).

Indeed, a series of studies have demonstrated that long-term music expertise renders observers more sensitive to temporal misalignments [63-65, 81]. For instance, Lee and Noppeney (2011) [4] showed that amateur pianists had a narrower temporal binding window than naive observers specifically for piano music but not for speech stimuli, where all observers have comparable expertise. Further, conductors as compared to musicians performed better at a task that requires synchronizing to audiovisual point-light representation of six single beat gestures, whereas musicians and non-musicians did not differ in terms of their synchronization abilities [82]. A more recent study provides some initial tentative evidence that observer's temporal sensitivity depend not only on music training per se but also to some extent on the specific music instrument the observer practiced [83]. More specifically, pianists showed a heightened sensitivity to audiovisual temporal misalignments selectively for piano rather than clarinet or violin music (see experiment 2 [83]). Yet, the interaction between music instrument and instrument training was not significant. Further psychophysical studies are needed to investigate the specificity and generalization of music training on the perception of audiovisual signals across different instruments.

Long-term music training frequently involves learning of new notations, i.e. a system that establishes a new symbolic mapping between sounds and visual symbols (see Table 1i.b and 1i.c for a summary). This provides the opportunity to investigate the effect of music training on processing of audiovisual incongruencies at a symbolic level. Several electrophysiological studies have demonstrated that musicians as compared to non-musicians exhibit different scalp topography and functional connectivity for oddball detection using audiovisual symbolic music stimuli [84-87] and this generalizes to audio-tactile stimulation [85, 88]. As shown in a series of MEG studies, this sensitivity to audiovisual incongruencies can also be learnt via short-term perceptual training [85, 89] (see Table 1ii. for a summary).

Collectively, the results suggest that internal forward models that are fine-tuned to a particular action such as piano playing may provide a supplementary mechanism for making predictions about the relative timing of audiovisual signal components. In line with this conjecture, a recent neuroimaging study [4] revealed activation increases for audiovisual asynchronous relative to synchronous music and speech stimuli not only in low level audiovisual regions and the superior temporal sulci that have previously been revealed for simple and complex audiovisual stimuli [90-92], but also in the cerebellar-premotor circuitry that is thought to instantiate a forward model in motor control. Importantly, the asynchrony responses indexing a prediction error were increased for piano players relative to non-musicians selectively for piano music but not for speech. Moreover, the premotor asynchrony effects predicted musicians' perceptual sensitivity to audiovisual asynchrony for piano music [4, 93]. Collectively, these studies suggest audiovisual temporal binding in perception recruits neural processes related to action production and observation. In addition to its well-

established effects on auditory processing and wider cognition [79, 94-96], music practice refines an action-specific internal forward model as a supplementary mechanism that enables more precise predictions of the relative timings of the auditory and visual signals. This line of research highlights intimate links between sensori-motor experience and audiovisual perception, whereby everyday interactions with the environment determine whether and how human observers integrate auditory and visual inputs into a unified percept (Fig. 3).

Fig. 3 about here

Conclusions

Audiovisual scene analysis requires the brain to infer the world's causal structure. For music and speech perception temporal regularities are critical cues informing the brain whether signals are caused by common sources and should be integrated into a unified percept. In line with models of predictive coding, we suggest that the brain fine-tunes an internal model to the statistical signal regularities of the environment. This internal model generates cross-sensory and sensori-motor temporal predictions as a mechanism to arbitrate between integration and segregation of signals from different senses.

Acknowledgment

This research was funded by the ERC (multsens). We thank Samuel Jones for comments on a previous version of this manuscript.

References

1. Ganesh, A.C., F. Berthommier & J.L. Schwartz. 2016. Audio Visual Integration with Competing Sources in the Framework of Audio Visual Speech Scene Analysis. *Adv. Exp. Med. Biol.* **894**: 399-408.
2. Gau, R. & U. Noppeney. 2016. How prior expectations shape multisensory perception. *Neuroimage*. **124**: 876-886.
3. Kording, K.P., U. Beierholm, W.J. Ma*, et al.* 2007. Causal inference in multisensory perception. *PLoS One*. **2**: e943.
4. Lee, H. & U. Noppeney. 2011. Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proc. Natl. Acad. Sci. USA*. **108**: E1441-1450.
5. Magnotti, J.F., W.J. Ma & M.S. Beauchamp. 2013. Causal inference of asynchronous audiovisual speech. *Front. Psychol.* **4**: 798.
6. Munhall, K.G., P. Gribble, L. Sacco*, et al.* 1996. Temporal constraints on the McGurk effect. *Percept. Psychophys.* **58**: 351-362.
7. Nahorna, O., F. Berthommier & J.L. Schwartz. 2015. Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *J. Acoust. Soc. Am.* **137**: 362-377.
8. van Wassenhove, V., K.W. Grant & D. Poeppel. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*. **45**: 598-607.
9. Bertelson, P. & M. Radeau. 1981. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Percept. Psychophys.* **29**: 578-584.
10. Bonath, B., T. Noesselt, K. Krauel*, et al.* 2014. Audio-visual synchrony modulates the ventriloquist illusion and its neural/spatial representation in the auditory cortex. *Neuroimage*. **98**: 425-434.
11. Rohe, T. & U. Noppeney. 2016. Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices. *Curr. Biol.* **26**: 509-514.
12. Wallace, M.T., G.E. Roberson, W.D. Hairston*, et al.* 2004. Unifying multisensory signals across time and space. *Exp. Brain Res.* **158**: 252-258.
13. Parise, C.V. & M.O. Ernst. 2016. Correlation detection as a general mechanism for multisensory integration. *Nat. Commun.* **7**: 11543.
14. Parise, C.V., C. Spence & M.O. Ernst. 2012. When correlation implies causation in multisensory integration. *Curr. Biol.* **22**: 46-49.
15. Knill, D.C. & A. Pouget. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**: 712-719.
16. Rohe, T. & U. Noppeney. 2015. Sensory reliability shapes perceptual inference via two mechanisms. *J. Vis.* **15**: 22.
17. Shams, L. & U.R. Beierholm. 2010. Causal inference in perception. *Trends Cogn. Sci.* **14**: 425-432.
18. Wozny, D.R., U.R. Beierholm & L. Shams. 2010. Probability matching as a computational strategy used in perception. *PLoS Comput. Biol.* **6**.
19. Alais, D. & D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* **14**: 257-262.
20. Ernst, M.O. & M.S. Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*. **415**: 429-433.

21. Helbig, H.B., M.O. Ernst, E. Ricciardi*, et al.* 2012. The neural mechanisms of reliability weighted integration of shape information from vision and touch. *Neuroimage*. **60**: 1063-1072.
22. Hillis, J.M., S.J. Watt, M.S. Landy*, et al.* 2004. Slant from texture and disparity cues: optimal cue combination. *J. Vis.* **4**: 967-992.
23. Jacobs, R.A. 1999. Optimal integration of texture and motion cues to depth. *Vision Res.* **39**: 3621-3629.
24. Knill, D.C. & J.A. Saunders. 2003. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res.* **43**: 2539-2558.
25. Rohe, T. & U. Noppeney. 2015. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol.* **13**: e1002073.
26. McGurk, H. & J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*. **264**: 746-748.
27. Magnotti, J.F. & M.S. Beauchamp. 2017. A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS Comput. Biol.* **13**: e1005229.
28. Saldana, H.M. & L.D. Rosenblum. 1993. Visual influences on auditory pluck and bow judgments. *Percept. Psychophys.* **54**: 406-416.
29. Schutz, M. & M. Kubovy. 2009. Causality and cross-modal integration. *J Exp Psychol Hum Percept Perform*. **35**: 1791-1810.
30. Fiser, J., P. Berkes, G. Orban*, et al.* 2010. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14**: 119-130.
31. Shen, S. & W.J. Ma. 2016. A detailed comparison of optimality and simplicity in perceptual decision making. *Psychol. Rev.* **123**: 452-480.
32. Bastos, A.M., W.M. Usrey, R.A. Adams*, et al.* 2012. Canonical microcircuits for predictive coding. *Neuron*. **76**: 695-711.
33. Friston, K. 2010. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**: 127-138.
34. Friston, K. & S. Kiebel. 2009. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**: 1211-1221.
35. Gagnepain, P., R.N. Henson & M.H. Davis. 2012. Temporal predictive codes for spoken words in auditory cortex. *Curr. Biol.* **22**: 615-621.
36. Muckli, L., F. De Martino, L. Vizioli*, et al.* 2015. Contextual Feedback to Superficial Layers of V1. *Curr. Biol.* **25**: 2690-2695.
37. Sedley, W., P.E. Gander, S. Kumar*, et al.* 2016. Neural signatures of perceptual inference. *Elife*. **5**: e11476.
38. Sohoglu, E., J.E. Peelle, R.P. Carlyon*, et al.* 2012. Predictive top-down integration of prior knowledge during speech perception. *J. Neurosci.* **32**: 8443-8453.
39. Summerfield, C., T. Egner, M. Greene*, et al.* 2006. Predictive codes for forthcoming perception in the frontal cortex. *Science*. **314**: 1311-1314.
40. Tuennerhoff, J. & U. Noppeney. 2016. When sentences live up to your expectations. *Neuroimage*. **124**: 641-653.
41. Baart, M. & J. Vroomen. 2010. Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neurosci. Lett.* **471**: 100-103.
42. Blank, H. & K. von Kriegstein. 2013. Mechanisms of enhancing visual-speech recognition by prior auditory information. *Neuroimage*. **65**: 109-118.

43. Noppeney, U., O. Josephs, J. Hocking*, et al.* 2008. The effect of prior visual information on recognition of speech and sounds. *Cereb. Cortex*. **18**: 598-609.
44. Vroomen, J. & M. Baart. 2009. Recalibration of phonetic categories by lipread speech: measuring aftereffects after a 24-hour delay. *Lang. Speech*. **52**: 341-350.
45. Yuan, X., C. Bi, H. Yin*, et al.* 2014. The recalibration patterns of perceptual synchrony and multisensory integration after exposure to asynchronous speech. *Neurosci. Lett.* **569**: 148-152.
46. Hirsh, I.J. & C.E. Sherrick, Jr. 1961. Perceived order in different sense modalities. *J. Exp. Psychol.* **62**: 423-432.
47. Stone, J.V., N.M. Hunkin, J. Porrill*, et al.* 2001. When is now? Perception of simultaneity. *Proc. Biol. Sci.* **268**: 31-38.
48. Zampini, M., S. Guest, D.I. Shore*, et al.* 2005. Audio-visual simultaneity judgments. *Percept. Psychophys.* **67**: 531-544.
49. Zampini, M., D.I. Shore & C. Spence. 2003. Audiovisual temporal order judgments. *Exp. Brain Res.* **152**: 198-210.
50. Denison, R.N., J. Driver & C.C. Ruff. 2012. Temporal structure and complexity affect audio-visual correspondence detection. *Front. Psychol.* **3**: 619.
51. Vroomen, J. & M. Keetels. 2010. Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* **72**: 871-884.
52. Fujisaki, W., S. Shimojo, M. Kashino*, et al.* 2004. Recalibration of audiovisual simultaneity. *Nat. Neurosci.* **7**: 773-778.
53. Simon, D.M., J.P. Noel & M.T. Wallace. 2017. Event Related Potentials Index Rapid Recalibration to Audiovisual Temporal Asynchrony. *Front. Integr. Neurosci.* **11**: 8.
54. Chandrasekaran, C., A. Trubanova, S. Stillittano*, et al.* 2009. The natural statistics of audiovisual speech. *PLoS Comput. Biol.* **5**: e1000436.
55. Dixon, N.F. & L. Spitz. 1980. The detection of auditory visual desynchrony. *Perception*. **9**: 719-721.
56. van Wassenhove, V., K.W. Grant & D. Poeppel. 2005. Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. USA*. **102**: 1181-1186.
57. Maier, J.X., M. Di Luca & U. Noppeney. 2011. Audiovisual asynchrony detection in human speech. *J. Exp. Psychol. Hum. Percept. Perform.* **37**: 245-256.
58. Sanchez-Garcia, C., J.T. Enns & S. Soto-Faraco. 2013. Cross-modal prediction in speech depends on prior linguistic experience. *Exp. Brain Res.* **225**: 499-511.
59. Powers, A.R., 3rd, M.A. Hevey & M.T. Wallace. 2012. Neural correlates of multisensory perceptual learning. *J. Neurosci.* **32**: 6263-6274.
60. Powers, A.R., 3rd, A.R. Hillock & M.T. Wallace. 2009. Perceptual training narrows the temporal window of multisensory binding. *J. Neurosci.* **29**: 12265-12274.
61. Stevenson, R.A., M.M. Wilson, A.R. Powers*, et al.* 2013. The effects of visual training on multisensory temporal processing. *Exp. Brain Res.* **225**: 479-489.
62. Grant, K.W., V.v. Wassenhove & D. Poeppel. 2004. Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication*. **44**: 43-53.
63. Petrini, K., S. Dahl, D. Rocchesso*, et al.* 2009. Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Exp. Brain Res.* **198**: 339-352.
64. Petrini, K., S.P. Holt & F. Pollick. 2010. Expertise with multisensory events eliminates the effect of biological motion rotation on audiovisual synchrony perception. *J. Vis.* **10**: 2.

65. Petrini, K., M. Russell & F. Pollick. 2009. When knowing can replace seeing in audiovisual integration of actions. *Cognition*. **110**: 432-439.
66. Bastos, A.M., J. Vezoli, C.A. Bosman*, et al.* 2015. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*. **85**: 390-401.
67. Arnal, L.H., V. Wyart & A.L. Giraud. 2011. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. Neurosci.* **14**: 797-801.
68. Arnal, L.H. & A.L. Giraud. 2012. Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* **16**: 390-398.
69. Lee, H. & U. Noppeney. 2014. Temporal prediction errors in visual and auditory cortices. *Curr. Biol.* **24**: R309-310.
70. Chen, J.L., V.B. Penhune & R.J. Zatorre. 2008. Listening to musical rhythms recruits motor regions of the brain. *Cereb. Cortex*. **18**: 2844-2854.
71. Lahav, A., E. Saltzman & G. Schlaug. 2007. Action representation of sound: audiomotor recognition network while listening to newly acquired actions. *J. Neurosci.* **27**: 308-314.
72. Zatorre, R.J., J.L. Chen & V.B. Penhune. 2007. When the brain plays music: auditory-motor interactions in music perception and production. *Nat. Rev. Neurosci.* **8**: 547-558.
73. Morillon, B. & S. Baillet. 2017. Motor origin of temporal predictions in auditory attention. *Proc. Natl. Acad. Sci. USA*.
74. Wolpert, D.M., R.C. Miall & M. Kawato. 1998. Internal models in the cerebellum. *Trends Cogn. Sci.* **2**: 338-347.
75. Grahn, J.A. & J.B. Rowe. 2009. Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. *J. Neurosci.* **29**: 7540-7548.
76. Lewis, P.A. & R.C. Miall. 2003. Brain activation patterns during measurement of sub- and supra-second intervals. *Neuropsychologia*. **41**: 1583-1592.
77. O'Reilly, J.X., M.M. Mesulam & A.C. Nobre. 2008. The cerebellum predicts the timing of perceptual events. *J. Neurosci.* **28**: 2252-2260.
78. Penhune, V.B., R.J. Zattore & A.C. Evans. 1998. Cerebellar contributions to motor timing: a PET study of auditory and visual rhythm reproduction. *J. Cogn. Neurosci.* **10**: 752-765.
79. Kraus, N. & B. Chandrasekaran. 2010. Music training for the development of auditory skills. *Nat. Rev. Neurosci.* **11**: 599-605.
80. Munte, T.F., E. Altenmuller & L. Jancke. 2002. The musician's brain as a model of neuroplasticity. *Nat. Rev. Neurosci.* **3**: 473-478.
81. Lee, H. & U. Noppeney. 2014. Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Front. Psychol.* **5**: 868.
82. Luck, G. & S. Nte. 2007. An investigation of conductors' temporal gestures and conductor— musician synchronization, and a first experiment. *Psychology of Music*. **36**: 81-99.
83. Bishop, L. & W. Goebl. 2014. Context-specific effects of musical expertise on audiovisual integration. *Front. Psychol.* **5**: 1123.
84. Nichols, E.S. & J.A. Grahn. 2016. Neural correlates of audiovisual integration in music reading. *Neuropsychologia*. **91**: 199-210.
85. Pantev, C., E. Paraskevopoulos, A. Kuchenbuch*, et al.* 2015. Musical expertise is related to neuroplastic changes of multisensory nature within the auditory cortex. *Eur. J. Neurosci.* **41**: 709-717.

86. Paraskevopoulos, E., A. Kraneburg, S.C. Herholz*, et al.* 2015. Musical expertise is related to altered functional connectivity during audiovisual integration. *Proc. Natl. Acad. Sci. USA*. **112**: 12522-12527.
87. Paraskevopoulos, E., A. Kuchenbuch, S.C. Herholz*, et al.* 2012. Musical expertise induces audiovisual integration of abstract congruency rules. *J. Neurosci.* **32**: 18196-18203.
88. Landry, S.P. & F. Champoux. 2017. Musicians react faster and are better multisensory integrators. *Brain Cogn.* **111**: 156-162.
89. Paraskevopoulos, E., A. Kuchenbuch, S.C. Herholz*, et al.* 2014. Multisensory integration during short-term music reading training enhances both uni- and multisensory cortical processing. *J. Cogn. Neurosci.* **26**: 2224-2238.
90. Lewis, R. & U. Noppeney. 2010. Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *J. Neurosci.* **30**: 12329-12339.
91. Miller, L.M. & M. D'Esposito. 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* **25**: 5884-5893.
92. Noesselt, T., J.W. Rieger, M.A. Schoenfeld*, et al.* 2007. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* **27**: 11431-11441.
93. Lu, Y., E. Paraskevopoulos, S.C. Herholz*, et al.* 2014. Temporal processing of audiovisual stimuli is enhanced in musicians: evidence from magnetoencephalography (MEG). *PLoS One*. **9**: e90686.
94. Musacchia, G., M. Sams, E. Skoe*, et al.* 2007. Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proc. Natl. Acad. Sci. USA*. **104**: 15894-15898.
95. Patel, A.D. & J.R. Iversen. 2007. The linguistic benefits of musical abilities. *Trends Cogn. Sci.* **11**: 369-372.
96. Wong, P.C., E. Skoe, N.M. Russo*, et al.* 2007. Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* **10**: 420-422.
97. Petrini, K., F.E. Pollick, S. Dahl*, et al.* 2011. Action expertise reduces brain activity for audiovisual matching actions: an fMRI study with expert drummers. *Neuroimage*. **56**: 1480-1492.

Figure Legends

Figure 1

Bayesian Causal Inference model

The generative model of Bayesian Causal Inference determines whether the visual input 'sight of a violin' and the auditory input 'music melody' are generated by common (C=1) or independent (C=2) sources (for details see [3]). For a common source, the 'true' audiovisual property in question (e.g. location, timing etc.: S_{AV}) is drawn from one prior distribution. For independent sources, the 'true' auditory (S_A) and 'true' visual (S_V) properties are drawn independently from this prior distribution. We introduce independent sensory noise to generate auditory (X_A) and visual (X_V) inputs.

Figure 2

Predictive coding and temporal predictions across the senses

According to predictive coding, backward connections (white) provide predictions from higher to subordinate cortical levels. Conversely, forward connections (black) furnish the prediction error that is computed at each cortical level as the difference between top-down predictions and bottom-up inputs. Prior expectations based on lifelong exposure to natural speech and music stimuli molds an internal model that enables the brain to predict the temporal relationship of auditory and visual signals. For visual leading signal (as in the example illustrated in the figure), the preceding visual signal induces the brain to generate temporal predictions across auditory and visual cortices, so that the delayed auditory signal elicits a prediction error signal in

the auditory cortices (and vice versa auditory leading signals elicit prediction errors signals in the visual cortices).

Figure 3

Internal forward models for sensori-motor temporal predictions

Internal forward models map from the motor plan of the intended action (e.g. piano playing) onto its sensory consequences. They are learnt via feedback by minimizing the prediction error, i.e. the difference between the predicted sensory consequences and the actual sensory consequences (e.g. piano sounds, visible finger movements, tactile sensations) that are caused by the action.

Table Legends

Table 1

Table 1. Summary of research studies that examined the effects of (i) long-term music training on: a. perception of audiovisual naturalistic speech and music stimuli, b. perception of audiovisual symbolic music stimuli, and c. perception of audio-tactile symbolic music stimuli, and (ii) short-term music training on perception of audiovisual music.

Table 1. Summary of research studies that examined the effects of (i) long-term music training on: a. perception of audiovisual naturalistic speech and music stimuli, b. perception of audiovisual symbolic music stimuli, and c. perception of audio-tactile symbolic music stimuli, and (ii) short-term music training on perception of audiovisual music.

Studies (first author)	Type of Data	Participants	Type of stimuli	Task	Results
i. Long-term music training influences processing of audiovisual speech and/or music					
a. Perception of audiovisual naturalistic speech and music stimuli					
Bishop[83]	Behavioural	Expert musicians (clarinetists, pianists, violinists)	Duo performances of 3 pieces of music	AV-SJ task: 3 stimulus type X 9 AV delays (0, ±0.04, ±0.12, ±0.2, ±0.28 s)	Musicians were most sensitive to asynchrony for piano stimuli and least sensitive to asynchrony for violin stimuli. Size of TIW: violin > clarinet > piano stimuli. TIW decreased with increasing musical experience.
Lee[81]	Behavioural	Amateur pianists vs. non-musicians	AV speech syllables and sentences, AV sinewave speech syllables and sentences, AV music tones and music melodies	AV-SJ task: 3 stimulus type X 2 stimulus duration X 13 AV delays (-0.36:0.06:0.36 s)	Musicians > non-musicians: Narrower TIW for music and sinewave speech but not speech stimuli; TIW for music decreased with amount of piano practice
Lee[4]	Behavioural and fMRI	Amateur pianists vs. non-musicians	AV speech sentences and AV music melodies	AV-SJ task outside scanner: 2 stimulus type X 13 AV delays (-0.36:0.06:0.36 s); AV passive viewing task inside scanner: 2 stimulus type X 3 AV delays (0 ms, ±0.24 s)	Musicians > non-musicians: Narrower TIW for music but not speech stimuli; Increased activation in bilateral pSTS, left premotor and left cerebellar region
Musacchia[94]	Brainstem	Amateur	Visual: male speaker	Task: subjects were to	Musicians > non-musicians:

	EEG	musicians vs. non-musicians	articulating the syllable “da”, musician bowing a cello; Auditory: speech syllable “da”, musical sound of a cello being bowed (note G2)	silently count the number of target stimuli (slightly longer in duration than non-targets) they saw or heard and then report that number at the end of each block	earlier and larger brainstem responses for both speech and music stimuli presented in auditory and AV conditions.
b. Perception of audiovisual symbolic music stimuli					
Luck[82]	Behavioural	Conductors vs. musicians vs. non-musicians	AV point-light representations of six single-beat gestures (differ in terms of degree of curvature) produced by two conductors (1 novice, 1 experienced)	2 conductor type X 3 gestures type Task: participants pressed space bar “in synchrony” with the stimuli they were presented with.	Conductors synchronized more consistently than musicians, whereas musicians and non-musicians did not differ in their synchronization abilities.
Petrini[63]	Behavioural	Expert jazz drummers vs. novices	Experiment 1 and 2: AV point light displays of drumming actions	Experiment1: AV-SJ task: 3 tempos X 3 accents X 9 AV delays (-0.267:0.067:0.267); Experiment 2: AV-SJ task: 2 AV incongruent X 9 AV delays (-0.267:0.067:0.267)	Musicians > non-musicians: Narrower TIW Non-musicians, but not musicians, showed increased sensitivity to asynchrony as a function of tempo.
Petrini[65]	Behavioural	Expert jazz drummers vs. novices	AV point light displays of drumming actions	AV-SJ task: 9 AV delays (-0.267:0.067:0.267) X 2 visual displays (elimination or inclusion of drumstick-drumhead impact point)	Non-musicians > musicians: unable to detect asynchrony when there is no information of the drumstick-drumhead impact point. For condition with drumstick-drumhead impact point information, musicians perceived best AV alignment when sight preceded sound, whereas for condition without drumstick-drumhead impact point information, musicians perceived best AV alignment when sound co-occurred or preceded with sight.

Petrini[64]	Behavioural	Expert jazz drummers vs. novices	AV point light displays of drumming actions	AV-SJ and TOJ tasks: 9 AV delays (-0.267:0.067:0.267) X 4 orientation views by participants	Musicians > non-musicians: narrower TIW for SJ task but not for TOJ task. More sensitivity to asynchrony. Non-musicians, but not musicians, showed less sensitivity to asynchrony when orientation view became less natural for SJ-task.
Nichols[84]	EEG	Amateur musicians vs. non-musicians	Visual: two musical notes on a treble-clef staff; Auditory: two 300 ms pure tones	Oddball paradigm. 85% of trials were AV congruent trials, 15% of trials were AV incongruent trials (either the visual stimulus deviated from the standard, the auditory stimulus deviated from the standard or both auditory and visual stimuli deviated from the standard).	Non-musicians > musicians: different scalp topography and more negative amplitudes at electrode Cz. Amplitudes to congruent stimuli were less negative as training increased, but amplitudes to incongruent stimuli did not change with training. Musicians > non-musicians: larger P300 and different scalp topography.
Pantev[85]	MEG	Expert musicians vs. non-musicians	Experiment 1: Visual: simplified music reading modus representing the pitch height of each tone (the higher the tone, the higher the position of the circle); Auditory: 5 tone melodies, AV incongruency (violation of the rule), auditory mismatch (timbre) or a visual mismatch (color) Experiment 3: see [93]	Experiment 1: oddball paradigm Task: participants indicated if the AV stimuli were congruent or incongruent and if there was a tone sounding differently from all others, or if a disk was of a different color. Experiment 3: 4 AV delays (0, +0.15, +0.2, +0.25 s) Experiment 3: see [93]	Experiment 1: Musicians > non-musicians: increased difference for incongruent > congruent trials in right auditory cortex Experiment 3: Musicians > non-musicians: increased difference for synchronous > asynchronous trials in left auditory cortex
Paraskevopoulos[87]	MEG	Expert musicians vs.	Visual: simplified music reading modus representing the pitch height of each tone	Oddball detection. Task: participants indicated	Musicians > non-musicians: increased activity in frontal, temporal, and occipital regions

		non-musicians	(the higher the tone, the higher the position of the circle); Auditory: 5 tone melodies, AV incongruency (violation of the rule), auditory mismatch (timbre) or a visual mismatch (color)	if the AV stimuli were congruent or incongruent and if there was a tone sounding differently from all others, or if a disk was of a different color.	as a response to AV incongruency, unisensory auditory and visual mismatch responses.
Paraskevopoulos [86]	MEG	Expert musicians vs. non-musicians	Visual: simplified music reading modus representing the pitch height of each tone (the higher the tone, the higher the position of the circle); Auditory: 5 tone melodies, AV incongruency (violation of the rule)	Congruent (the higher the pitch, the higher the position) and incongruent AV pairings.	Musicians and non-musicians showed different connectivity patterns for integration AV information.
Lu[93]	Behavioural and fMRI	Expert musicians vs. non-musicians	Visual: black circular dot; Auditory: 200 ms of sinusoidal tone of 880 Hz;	4 AV delays (0, +0.15, +0.2, +0.25 s) Task: participants judged if the stimulus presented was a synchronous, asynchronous or control trial.	Musicians > non-musicians: more accurate when judging if the AV stimuli were synchronous or asynchronous; Increased activity in left pSTS, insula and post-central gyrus for synchronous trials, and increased activity in left cerebellum for asynchronous trials.
Petrini[97]	fMRI	Expert jazz drummers vs. novices	AV point light displays of drumming actions	Experiment 1: AV synchronous and asynchronous (visual leading) stimuli; Experiment 2: AV synchronous congruent vs. AV synchronous incongruent stimuli	Experiment 1: Musicians > non-musicians: more sensitive to AV asynchrony; reduced activity in bilateral cerebellum and left parahippocampal gyrus. Experiment 2: Musicians showed increased activity for incongruent > congruent AV synchronous stimuli in right IPL, right ITG, right MFG and right precentral gyrus. Non-musicians showed increased

					activity for incongruent > congruent AV synchronous stimuli in right ITG.
c. Perception of audio-tactile symbolic music stimuli					
Landry[88]	Behavioural	Expert musicians vs. non-musicians	Tactile: 50ms of vibration of 200 Hz presented by a vibrotactile device; Auditory: 50 ms of white noise burst	Task: Participants indicated immediately upon the perception of an auditory, tactile or synchronous audio-tactile stimulation.	Musicians > non-musicians: faster when responding to an audio-tactile stimulation.
Pantev[85]	MEG	Expert musicians vs. non-musicians	Experiment 2: Tactile stimulation of left hand (index, middle, ring and litter fingers); Auditory: 4 possible pitches, starting with the lowest tone corresponding to the stimulation of the index finger, second lowest tone to the middle finger, second highest tone to the ring finger and highest tone to the little finger, an audio-tactile incongruency, an auditory mismatch (timbre) or a tactile mismatch (change of location of the tactile stimulation).	Experiment 2: oddball paradigm Task: participants indicated if the audio-tactile stimuli were congruent or incongruent and if there was a tone sounding differently from all others, or if a disk was of a different color.	Experiment 2: Musicians > non-musicians: increased difference for incongruent > congruent trials in left auditory cortex.
ii. Short-term music training influences processing of audiovisual music					
Pantev[85]	MEG	Non-musicians	Visual: simplified music reading modus representing the pitch height of each tone (the higher the tone, the higher the position of the circle); Auditory: 5 tone melodies, audiovisual incongruency (violation of the rule), auditory mismatch	2 training group (group trained with audiovisual stimulus (AV-Int) or group with separate auditory and visual training (AV-Sep)); same task as Experiment 1; 5 training sessions in 1 week; MEG recorded after 1st training and the last	AV-Int: Incongruent > congruent trials in left auditory cortex, whereas AV-Sep: no effect in AV integration

			(timbre) or a visual mismatch (color)	immediately before the second MEG recording.	
Paraskevopoulos [89]	MEG	Non-musicians	Experiment 1: Visual: simplified music reading modus representing the pitch height of each tone (the higher the tone, the higher the position of the circle); Auditory: 5 tone melodies, AV incongruency (violation of the rule), auditory mismatch (timbre) or a visual mismatch (color)	2 training group (group trained with audiovisual stimulus (AV-int) or group with separate auditory and visual training (AV-Sep)); same task as Experiment 1; 5 training sessions in 1 week: 1st training immediately after MEG recording and 5th training immediately before MEG recording. MEG recorded before training (Pre) and immediately after last training (Post).	Post > Pre: better at detecting incongruent AV trials. Post > Pre & AV-Int > AV-Sep & incongruent > congruent AV trials: left superior frontal gyrus and left STG

Legend: AV = audiovisual; MEG = Magnetoencephalography; EEG = electroencephalography; fMRI = functional magnetic resonance imaging; SJ = synchrony judgment; TOJ = temporal order judgment; TIW = temporal integration window; pSTS = posterior superior temporal sulcus; STG = superior temporal gyrus; IPL = inferior parietal lobule; ITG = inferior temporal gyrus; MFG = middle frontal gyrus; positive AV delays indicate that visual precedes auditory signal, while negative AV delays indicate that visual lags auditory signal.