In R. Brown (ed.), *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, *Springer*, 2014.

Making Sense of Multiple Senses

By Kevin Connolly

Abstract: In the case of ventriloquism, seeing the movement of the ventriloquist dummy's mouth changes your experience of the auditory location of the vocals. Some have argued that cases like ventriloquism provide evidence for the view that at least some of the content of perception is fundamentally multimodal. In the ventriloquism case, this would mean your experience has constitutively audio-visual content (not just a conjunction of an audio content and visual content). In this paper, I argue that cases like ventriloquism do not in fact warrant that conclusion. I then try to make sense of crossmodal cases without appealing to fundamentally multimodal content.

1. Introduction

In the *McGurk effect*, a subject views a video of a person saying one set of syllables (e.g. *ga-ga*), while the audio has been redubbed to a second set of syllables (e.g., *ba-ba*). The subject experiences yet a third set of syllables, distinct from the first two sets (e.g., *da-da*) (McGurk and MacDonald, 1976, p. 747). The McGurk effect is a crossmodal experience. Crossmodal experiences are a kind of multimodal experience, that is, a kind of experience that involves more than one sense modality. More precisely put, a crossmodal experience is a kind of multimodal experience where an input in one sense modality changes what you experience in another sense modality. In the McGurk effect, for instance, the visual input of seeing the person mouth *ga-ga* changes the auditory input (*ba-ba*) to what you in fact hear (*da-da*).

Tim Bayne (forthcoming) has recently proposed two different interpretations of crossmodal cases such as the McGurk effect. On a strictly causal interpretation, seeing the person mouth *ga-ga* causes you to hear *da-da* instead of *ba-ba*. According to this interpretation, integration occurs between processing in the auditory system and the visual system (more on this process later), but the result of that processing can be fully decomposed into an audio component and a visual component. So, while the processing is multisensory, the content of that processing

is not intrinsically multisensory. On a *constitutive* interpretation, on the other hand, the *ga-ga* visual input and *ba-ba* auditory input give you an experience that has constitutively audio and visual content (not just a conjunction of audio and visual content). According to this interpretation, the perceptual state that results from the processing cannot be fully decomposed into two unisensory token states, one auditory state and one visual.

Should we hold a constitutive or causal interpretation of crossmodal cases like the McGurk effect? This question can be re-formulated in the following way: in crossmodal cases, are *constitutively multimodal properties* part of your phenomenal content? There are several ways to understand what it means to be a constitutively multimodal property, and later in the paper, I examine some of these options. To start, one option (very roughly) is to hold that a multimodal property is something over and above the properties contributed by each of the sense modalities involved. In this way, a constitutively audio-visual property would be modeled on flavor properties—properties that are arguably not just the conjunction of the properties contributed by each of the sense modalities involved in flavor perception (taste, touch, and retronasal smell). Like flavor properties, multimodal properties might be defined relative to subjects of experience, or they could be defined as objective kinds (see Smith, 2013, for a discussion of this issue for flavors).

What does it mean for a multimodal property to be part of your *phenomenal content*. "Phenomenal content," I will hold, is "that component of a state's representational content which supervenes on its phenomenal character" (Bayne, 2009, pp. 386-387)? In a McGurk effect case, for instance, the question is whether there is a constitutively audio-visual property in your phenomenal content, or whether it is just an audio property plus a visual property in your phenomenal content.

We can interpret other crossmodal cases constitutively or causally as well. In the motion-bounce illusion, subjects look at a computer display of two disks moving steadily towards each other until they meet. If the subject hears a sound at or around the point of convergence, the disks typically appear to collide and bounce off one another. If the subject does not hear a sound, the disks appear to cross through one another (Sekuler et al., 1997). According to a strictly causal interpretation, the motion-bounce illusion is a case where the sound simply causes you to have a certain visual experience (given the right visual input). According to a constitutive interpretation, on the other hand, it is a case where you have a constitutively audio-visual experience.

Whether we take a constitutive or causal interpretation of crossmodal cases seems to determine, at least at first glance, whether we hold that some of the content of perception is fundamentally multimodal. If we hold a constitutive interpretation of the McGurk effect, for instance, then we hold that at least some of the content of perception is audio-visual. A strictly causal interpretation, on the other hand, does not commit us to that.[1]

In what follows, I argue against various reasons for thinking that content of crossmodal experiences is fundamentally multimodal. In the next three sections, I examine three different reasons one might hold that view, and I argue that none of them actually entail fundamentally multimodal content. I close by trying to make sense of crossmodal cases without appealing to fundamentally multimodal content.

2. Is Crossmodal Perception like Flavor Perception?

The constitutive interpretation of crossmodal cases comes in several different varieties. One variety (the weakest, in my view) models the constitutive interpretation after flavor

[1] I owe the basic point behind this paragraph to Susanna Siegel, who made the point at *The Unity of Consciousness and Sensory Integration* Conference at Brown University in November of 2011. In the subsequent discussion, Tim Bayne said he held the constitutive interpretation.

perception, or, at least, one understanding of flavor perception. Such a view is mentioned, although not endorsed, by Fiona Macpherson (2011, p. 449).

Flavor perception is not the product of a single sense. Rather, it arises from the combination of multiple sense modalities, including taste, touch, and retronasal smell (smell directed internally at the food you have just eaten, rather than at external objects) (Smith, 2013). For instance, if you plug your nose entirely while eating an orange, you will not be able to detect the flavor of the orange. This is because the sense of smell is necessary for experiencing the flavor. Without it, there is no flavor experience. Flavor experience arises only through the combination of smell, touch, and taste.

On this interpretation of flavor perception, when a particular flavor perception integrates the properties detected by taste, smell, and touch, it creates a new whole: a flavor property. Fiona Macpherson describes what an account of crossmodal cases would sound like if such cases were modeled after flavor perception:

> [W]e can imagine a case where the new information produced was such that it was none of the above—it could not be produced by a single sensory modality, it did not involve cross-modal content of a binding or other kind—it simply consisted of some brand new content. An example of such a case would be one account of flavour experiences. (2011, p. 449)

If the content in crossmodal cases were like the content of flavor perception, then the content would not simply be the sum of the contents of each of the individual sense modalities involved (like the contents of taste, touch, and retronasal smell in flavor perception), but rather something over and above those contents (like flavors in flavor perception). So, on this way of construing the constitutive view, the content of an experience of the McGurk effect is not just an audio content plus a visual content, but a single, new, audio-visual content.

Are such audio-visual properties part of the content of perception? Consider two other properties first: the property of *being a wren* and the property of *being red*. Even for someone with excellent discrimination, there might be fake wrens that are visually identical to real wrens when examined across all the same lighting conditions and angles. Arguably, this suggests that *being a wren* is not a perceptual property at all. The same conclusion does not follow for properties like colors. There is no such thing as a fake red that is visually identical to an authentic red. The idea is that, for red, if you duplicate its appearance properties, you duplicate the property. On the other hand, there can be visually indistinguishable fake wrens or robot wrens. For a property like *being a wren*, you can duplicate its appearance properties without duplicating the property. Michael Tye registers the same sort of principle for denying that properties are part of the perceptual content:

> It seems plausible to suppose that the property of being a tiger is not itself a feature represented by the outputs of the sensory modules associated with vision. Our sensory states do not track *this* feature. There might conceivably be creatures other than tigers that look to us phenomenally just like tigers. (1995, p. 141)

On Tye's view, the property of being a tiger is not likely to be represented in vision because you could duplicate every single one of its visual features, and still not duplicate the property of *being a tiger*.

Are some of the contents of perception fused multimodal units (fused audio-visual units, for instance)? I think that the answer is no, and one reason why is grounded in the test just described. Call *Q1*, your experience of the familiar ventriloquist and dummy routine, where you hear the sound of the ventriloquists voice as coming from the dummy's mouth, even though it is actually coming from the ventriloquist's lips. Call *Q2*, an experience of a ventriloquism fakery. The ventriloquist, it turns out, is a fraud, and so he has recorded himself and has placed a speaker playing the recording in the dummy's mouth. Now consider the plausible assumption that *Q1*

and *Q2* are phenomenally identical experiences: what it's like in *Q1* is exactly what it's like in *Q2*. But quite plausibly *Q*2 represents just a regular auditory property and a visual property, rather than a fused audio-visual property. If that's right, however, we need not hold that the content of *Q1* involves a fused audio-visual property, since we can explain that phenomenal type in terms of an auditory property and a visual property.

We can arrange the same sort of scenario for the McGurk effect. Call *R1* a particular McGurk effect experience: the experience of a subject who views a video of a person saying *ga-ga*, while the audio has been redubbed *ba-ba*, so that the subject experiences *da-da.* Call *R2*, an experience of a fake McGurk effect. *R2* is the experience of a subject who views a video of a person saying *ga-ga*, while the audio has been redubbed to *da-da* (Note that when this scenario was tested in MacDonald and McGurk, 1978, subjects heard *da-da* one-hundred percent of the time). Now consider the plausible claim that *R1* and *R2* are phenomenally identical experiences. Quite plausibly *R*2 just represents an auditory property (of a person saying *da-da*) and a visual property (of a person saying *ga-ga*), rather than a fused audio-visual property. But then we need not hold that the content of *Q1* involves a fused audio-visual property, since we can explain that phenomenal type in terms of an auditory property and a visual property.

Why think that the above cases should be explained as a conjunction of audio content and visual content, rather than as involving fused audio-visual content? One reason is that everyone agrees that audio and visual properties are represented in perception. Unlike fused audio-visual properties, audio and visual properties are uncontroversial candidates for the content of perception. The question is whether fused audio-visual properties are represented in addition to audio and visual properties, not instead of them. If we reject fused audio-visual content, and

appeal instead to audio content and visual content, our account of content is also more economical, since we don't need to posit a new kind of property.

Consider another reason for why fundamentally multimodal properties should not be modeled on flavor properties. In the founding study of the McGurk effect, the authors wrote, "A 'fused' response is one where information from the two modalities is transformed into something new with an element not presented in either modality…" (McGurk and MacDonald, 1976, p. 747). Note the sense in which the information is transformed into something new. When a subject experiences the McGurk effect and hears *da-da*, this is a new property in the sense that it is neither the input of the auditory system, nor the input of the visual system. But it is not new in another sense: it *can* be the input of the auditory system, and it *can* be the input of the visual system. Those systems can detect that property. On the other hand, the fusion involved in flavor perception is new in a different sense. It cannot be the input of any of the systems involved (taste, touch, or retronasal smell), since those systems cannot detect flavor properties by themselves. In short, the kind of fusion involved in flavor perception does not occur in crossmodal perception.

In the motion-bounce illusion, the crossmodal influence of the sound serves to modulate the particular motion that you see (you see one motion rather than another). But, of course, in a different context you could have seen that motion. It is a new property in the sense that it is not the input of the visual system in the motion-bounce scenario. But it is not new in another sense: it *can* be the input of the visual system. You do not need crossmodal influence to see the motion that you see. In the ventriloquist effect, the sense of vision influences audition. If you are blindfolded as you enter a movie theater, you will hear the sounds of the movie as coming from the sides of the theater. When you are finally unblindfolded, vision influences your audition.

Before, you heard the sounds as coming from the sides of the theater. Afterwards, you hear the sounds as coming from the screen. But you could already detect auditory location. The crossmodal influence serves to modulate the auditory location that you experience, as you perceive a new location for the sound. In the McGurk effect, vision influences audition. If you were to cover your ears and then uncover them while watching the video, your visual experience would not change. On the other hand, if you were to cover your eyes and then uncover them, you would hear different syllables in the two experiences. Your auditory perception changes after you see the person's lips move. You see a person saying one set of syllables, while the audio has been changed to a second set of syllables, but you experience yet a third set of syllables. But again, you could already hear syllables and see someone saying them. The crossmodal influence serves to modulate the syllables that you hear (you hear different syllables before and after you uncover your eyes).

3. Do We Perceive Audio-Visual Bounces?

In the motion-bounce illusion, audition influences vision. At first, you see the disks passing through one another. Your visual perception of the disk trajectories changes only after the introduction of a sound, and then you see them as colliding with one another. According to a constitutive interpretation of crossmodal cases, it is a case where you have a constitutively audio-visual experience. One variety of such an interpretation is to hold that *being a bounce* is part of the content, where that property is construed as an audio-visual property. What does it mean to be an audio-visual bounce? Matthew Nudds writes, "We often see something happen and hear a sound, and we perceive the sound to have been produced by what we saw happen, we experience the production of the sound" (2001, p. 218). We might construe the "bounce" in the motion-bounce illusion similarly. The idea is that we see the collision and rebound and hear the sound,

and we perceive the sound to be produced by the collision, thereby experiencing the production of the sound. The collision causes the sound in an audio-visual bounce.

Nudds defends the view that we experience the production of sound (as in the audio-visual bounce case) by arguing for the more general claim that we can perceive one event causing another. To this end, he claims that we can perceive scrapes, pushes, squashes, and so on (2001, p. 218). Nudds backs up this claim by saying, "For as long as we allow that people possess and use such concepts [like scrapes, pushes, squashes, etc.] and can apply them to things on the basis of perceiving the interactions between, then we should allow that causality, in this sense, can be perceived" (220). Of course, Nudds is right that no one denies that we possess and correctly apply such concepts. But it doesn't follow that those concepts actually pick out scrapes, pushes, squashes, etc. *as perceptible properties*. Plausibly, like many robust concepts, such as the concept EMPTY GAS TANK, we do not apply them based solely on a perception. Rather, we apply them based on a perception and a background belief. If the concepts SCRAPE, PUSH, and SQUASH are like the concept EMPTY GAS TANK in this way, then while we may possess and correctly apply such concepts, it does not follow that scrapes, pushes, and squashes can be perceived.

In the motion-bounce illusion, it might seem at first glance that your perception represents an audio-visual bounce. My claim is that that does not follow, at least from Nudds' considerations. His argument does not actually show that we can perceive one event causing another, so it does not provide a defense of the claim that we experience the production of sound (as in the audio-visual bounce case). Still, there is something right in what Nudds says: we need to think of crossmodal cases like the motion-bounce illusion as events, if we are to understand them. I explore this idea in the next section.

4. Do We Need Multimodal Content to Explain Multisensory Integration?

Crossmodal influence modulates properties for a particular purpose, namely, to reconcile them with the properties in another modality (Matthen et al., 2011). That is to say, crossmodal cases involve multisensory integration: "the brain's ability to synthesize the information that it derives from two or more senses" (Stein et al., 2002, p. 227). But why exactly do the inputs in crossmodal experience require integration or reconciliation? Why do the properties represented by one modality have to align with the properties represented by another at all?

As Casey O'Callaghan points out, "[G]iven divergent auditory and visual stimulation, it only makes sense to attempt in a principled manner to reconcile them if they are assumed to share a common source or cause. Otherwise, the notion that there is a conflict that requires resolution is unintelligible" (2008, p. 326). The idea is that in a crossmodal case, the inputs in two different modalities conflict because they are predicated of a common source or cause (whether it be an individual, object, or event). This conflict requires the reconciliation between the inputs, and what we experience is the product of that reconciliation.

I agree with O'Callaghan's claim that in crossmodal cases, the inputs in two different modalities conflict because they are predicated of a common source or cause (whether it be an individual, object, or event). But my claim is that if O'Callaghan's argument is properly understood, it does not entail that those individuals, objects, or events have multimodal content. Roughly and briefly, this is because O'Callaghan's argument is meant only to undermine the view that the content of perception can be exhausted by unimodal content. But such an argument does not compel us to accept multimodal content. This is because the non-unimodal content could be amodal content (that is, modality-independent content—content not shared by the senses, but rather content that outstrips the senses).

Suppose that for the ventriloquist effect, the motion-bounce illusion, and the McGurk effect you did not experience a crossmodal effect. For instance, suppose that you sit down in a movie theater and see people talking on the screen, and cars exploding, but you hear all of the sounds coming from the sides of the movie theater. It is a very unusual experience to see lips moving and hear a sound consistent with those movements, but coming from a different direction. One way to render the data consistent would be to realize the way that a sound system is set up in a movie theater. Instead of this, your sensory system reconciles the auditory and visual inputs for you. You hear the sounds as coming from the screen (although they are coming from the side of the theater).

To take another example, suppose that in the motion-bounce scenario, you simply heard a random sound when the disks intersected, and experienced the disks as crossing through each other rather than bouncing. Once again, that data would require reconciliation. Why was there a random sound? As with the ventriloquist effect, in the motion-bounce illusion, your sensory system reconciles the data. You see the disks as colliding with one another. The sound is heard as the sound of a collision. This makes sense of the random sound.

Suppose that in the McGurk scenario, you saw someone mouthing the syllables *ga-ga*, but heard someone repeating the syllables *ba-ba*. That data would require reconciliation. Typically you hear the syllables that you see a person mouthing, not some other syllables. Seeing someone mouth *ga-ga* while hearing *ba-ba* requires reconciliation. In the McGurk effect, your sensory system performs that task. Importantly, however, even though you are looking at someone mouthing the syllables *ga-ga*, your sensory system does not reconcile that by having you hear the syllables *ga-ga*. Instead, you hear the syllables *da-da*. This might seem to suggest

that the auditory and visual inputs are left unreconciled. But McGurk and MacDonald suggest an alternative hypothesis:

> [I]n a ba-voice/ga-lips presentation, there is visual information for [ga] and [da] and auditory information with features common to [da] and [ba]. By responding to the common information in both modalities, a subject would arrive at the unifying percept [da] (1976, p. 747).

When you hear *da-da*, McGurk and MacDonald suggest, this is not a failure to reconcile the ba-voice and the ga-lips. Rather, the ba-voice actually contains some informational features of the sound *da-da*, while the ga-lips contain some informational features of seeing someone say *da-da*. When you hear *da-da*, McGurk and MacDonald claim, you are reconciling auditory and visual data through their common informational features (I explain this further in the next section).

In crossmodal cases, the inputs in two different modalities conflict because they are predicated of a common source or cause (whether it be an individual, object, or event). It might seem at first glance that if we posit individuals, objects, or events as the common source or cause in crossmodal cases, we are positing multimodal content. O'Callaghan, however, is careful not to make that inference. Rather, he says, "[T]here is a dimension or component of perceptual content that must be characterized in multi-modal *or modality-independent terms*. This component either is shared by both vision and audition *or outstrips both the visual and the auditory*" (2008, p. 328, italics added for emphasis; see also pp. 327-332, and O'Callaghan, forthcoming, section 5.2). O'Callaghan's point is that we can construe the individuals, objects, or events in two different ways: *either* as both the content of modality one (e.g., audio content) and the content of modality two (e.g., visual content) *or* as neither the content of modality one nor the content of modality two but as content that outstrips them both. If we characterize the individuals, objects, or events in the second way, that is, in modality-independent terms, then we are not positing multi-modal content. We are positing amodal content.

Let's return to the two rival interpretations of crossmodal cases from the introduction of the paper. The idea was that we can take either a constitutive or causal interpretation of crossmodal cases, and that that determines whether we hold that the phenomenal contents of crossmodal experiences are constitutively multi-modal, or whether they are just unimodal. The assumption was that in a McGurk effect case, for instance, there is either a fundamentally multimodal audio-visual property in your phenomenal content or else just an audio property plus a visual property. But suppose that we hold, following O'Callaghan, that crossmodal cases require us to posit individuals, objects, or events so that we can make sense of why reconciliation needs to occur in the first place. Suppose also that we characterize those individuals, objects, and events in modality-independent terms. We then end up with a new position, one where the content of crossmodal cases is neither multimodal, nor simply unimodal, but rather amodal. The lesson is this: O'Callaghan's claim is that we need to posit some sort of common content, shared by different sense modalities, in order to explain why reconciliation needs to occur in crossmodal cases in the first place. But shared content does not entail multi-modal content.

O'Callaghan's main goal in his 2008 article is to argue against the view that unimodal content exhausts perceptual content. As he puts it:

> I wish to argue that understanding cases of cross-modal perception grounds an argument for the claim that there exist consciously accessible aspects of perceptual experience that are not unique or specific to a given experiential modality and that may be shared across modalities. The argument proceeds in two stages. The first aims to show that that there is a dimension or component of perceptual content that must be characterized in multi-modal or modality-independent terms. This component either is shared by both vision and audition or outstrips both the visual and the auditory. (p. 328)

Given that his goal is to argue against the view that unimodal content exhausts perceptual content, O'Callaghan seems satisfied to accept either multimodal or modality-independent (amodal) content, since both are non-unimodal content. At the same time, he clearly does

distinguish between the two options. Multimodal content is shared by, say, both vision and audition, while modality-independent (amodal) content outstrips both vision and audition.

5. Crossmodal Cases Without Fundamentally Multimodal Content

So far I have argued against various reasons for thinking that crossmodal cases show that at least some of the content of perception is fundamentally multimodal—that is, reasons for thinking that your experience has, say, constitutively audio-visual content (not just a conjunction of an audio content and visual content). I now want to try to make some sense of crossmodal cases without appealing to fundamentally multimodal content.

A 2004 study at Oxford's Crossmodal Research Lab showed that hearing an augmented sound of a crunch makes soft potato chips seem crisper and stale chips seem fresher (Zampini and Spence). In that study, a higher volume of a crunch sound correlated with the chips seeming crisper and fresher, while a lower volume correlated with the chips seeming softer and staler. The study showed that the sensory system is able to reconcile auditory data with gustatory data, in this case by modulating the experience of crispness or freshness.

Take a particular class of perceptible properties (the class of colors, or shapes, or sizes, or locations, or orientations, for instance), and for a substantial portion of its members *x*, *y*, and *z*, *x* is more similar to *y* than it is to *z*. For instance (as a first approximation), for colors, orange is more similar to red than it is to blue. For size, a peanut is more similar to a watermelon than it is to the Empire State Building. A more precise examination of similarity orderings shows that they are often multi-dimensional. Colors, for instance, are comparable along the dimensions of brightness, saturation, and hue (Matthen, 2005, p. 111). By utilizing those three dimensions, for a substantial portion of colors *x*, *y*, and *z*, *x* will be more similar to *y* than it is to *z*.

In what follows, I want to show how such a similarity structure might help us to understand crossmodal cases. For the class of crisp things, for instance, we can say of a substantial amount of its members that *x* is more similar in crispness to *y* than it is to *z*. In the Zampini and Spence study, as one's sensory system reconciles a flavor with a sound, the flavor appears more crisp or less crisp, more fresh or less fresh. In everyday situations (outside of the experimental context), when you hear a crunch sound of magnitude *x*, there would be a correlating magnitude of crispness *y*. In the experimental context, when you hear an augmented crunch sound of magnitude *x*, the actual magnitude of the crispness is less than *y*, but you perceive something more similar in magnitude to *y*.

Put another way, modality one detects a property (crunch volume) that can be located on a similarity space. Modality two detects a different property (crispness) that can be located on a similarity space. Certain points on each similarity space correlate with particular points on the other similarity space (crunchiness of magnitude X with crispness of magnitude Y, e.g.). A plausible story is that through learned experience, you build an association between the crunchiness of magnitude X and the crispness of magnitude Y. In crossmodal cases, each modality detects a particular property, one on each of the spaces (the crunchiness space and the crispness space), and these are properties that do not typically correlate. The crossmodal effect is to shift one of the properties, in experience, such that it is closer to its correlating point with the other experienced property. At bottom, this is just a shift along the continuum for a type of property that is already represented in perception. It is not any new kind of property.

My proposal is that hearing an augmented sound of a crunch can make stale potato chips seem crisper because crispness is a kind of property that *can* be reconciled with an aberrant crunch sound of magnitude *x*. Specifically, it can be made more similar to the magnitude of

crispness that typically corresponds with the magnitude of that sound. In the Zampini and Spence study, the same holds, *mutatis mutandis*, for the property of freshness.

But now consider our three crossmodal cases as cases that aim at data reconciliation. In the McGurk effect, as your sensory system reconciles a sound with a visual image, it modulates the sound. In most everyday situations (outside of the experimental context), when you see someone mouthing the syllables *ga-ga*, there would be a correlating sound: *ga-ga*. In the experimental context, when you see someone mouthing the syllables *ga-ga*, the actual sound is *ba-ba*, but you hear something more similar to *ga-ga*, namely, *da-da* (I will motivate the claim that these two sounds are more similar shortly).

We know from our own experience that some words sound more similar to each other than others. One piece of evidence for this is that we confuse some words with each other when we hear them, but do not confuse other words with each other. If we break down spoken words into their units, we can tell the same sort of story about these units, or *phonemes*. A phoneme *x* can sound more similar to a phoneme *y* than to another phoneme *z*. Todd M. Bailey and Ulrike Hahn have charted the similarity relations between phonemes in great detail (Bailey and Hahn, 2005; Hahn and Bailey, 2005). For instance, they argue that "/t/ is more similar to /d/ than to /l/" (where "/t/" represents a phoneme of t) (Bailey and Hahn, 2005, p. 339). According to them, this is why "tuck" sounds more similar to "duck" than it does to "luck." Phoneme similarity helps to explain why we sometimes confuse certain words when we hear them, but not others.

We need not commit to a single unified phoneme space, where each phoneme can be ordered in relation to every other phoneme (just as every color can be ordered in relation to every other color). Still, we can say that there are phoneme *spaces*. To use Bailey and Hahn's example, /t/ is more similar to /d/ than to /l/. My claim is that the McGurk effect exploits such spaces. *Da-*

da sounds more similar to *ga-ga* than *ba-ba* does. This account dovetails with McGurk and MacDonald's account of the McGurk effect. They speculate that "the acoustic waveform for [ba] contains features in common with that for [da] but not with [ga]…" (1976, p. 747). On their view, the similar acoustic waveform is what accounts for the similar sounds of *ba-ba* and *da-da*.

In the McGurk effect, the audio plays one sound (e.g., *ba-ba*), and the visual shows someone mouthing a second sound (e.g., *ga-ga*), but you hear yet a third sound (e.g., *da-da*). My suggestion is that your sensory system reconciles the aberrant sound (*ba-ba*) by making it more similar to the sound that typically would correspond with the image that you see (*ga-ga*). *Da-da* sounds more similar to *ga-ga* than *ba-ba* does.

According to McGurk and MacDonald, the ga-lips also contribute to data reconciliation in the McGurk effect (1976, p. 747). As I mentioned, they claim that the sound *ba-ba* shares some informational features in common with the sound *da-da* (they put this point in terms of a similar acoustic waveform). But they also claim that seeing someone say *ga-ga* shares some informational features with seeing someone say *da-da* (they cite the fact that lip movements for *ga-ga* are frequently misread as lip movements for *da-da*). According to their explanation, hearing *da-da* provides a unique solution to the conflicting visual and auditory data. It reconciles the auditory and visual data through their common informational features.

Typically, when you see someone mouthing "ga-ga," you hear the sound "ga-ga." Notice that in the McGurk effect, the association between seeing someone mouth "ga-ga" and hearing "ga-ga" is not strong enough to make someone hear "ga-ga." Instead you hear "da-da" when the audio is "ba-ba." Still, the weight of the association between seeing someone mouth "ga-ga" and hearing "ga-ga" is strong enough to shift the heard property from ba-ba (which is the input) along a perceptual dimension to da-da (which is what is heard). Why is the auditory pull from ba-ba to da-da, and not all the way to ga-ga? I think the similarity space makes sense of this.

Auditorily, ga-ga is more similar to da-da, than it is to ba-ba. The crossmodal effect is to shift the auditory property, in experience, such that it is closer to its correlating point with the other experienced property, the visual property. What you hear in the McGurk effect is more similar to the auditory correlate of what you see. Again, this is just a shift along the continuum for a type of property that is already represented in perception, rather than a new kind of property.

In the ventriloquist effect, as your sensory system reconciles an auditory location with what you see, it modulates the auditory location. Typically, when you see lips moving and hear a sound consistent with the lip movements, the location of that sound is the moving lips. In the ventriloquist effect, when you see the lip movements, the actual auditory location is from elsewhere, but you experience the location as from the moving lips. The ventriloquist effect operates on auditory location. In the ventriloquist effect, your sensory system reconciles an aberrant auditory location (e.g., the location of the sides of a movie theater) by making it more similar to the auditory location that typically would correspond with the image that you see (e.g., the movie screen).

Both the McGurk effect and the ventriloquist effect are cases where auditory and visual data conflict, and in both cases, vision is dominant. That is, in both cases, the auditory data reconciles with the visual data. Vision is not always dominant, however. In the motion-bounce illusion, for instance, as your sensory systems reconcile a visual image with what you hear, it modulates the visual image. Typically, when you see two objects coincide and hear a sound when they do, you see the motion we call "bouncing." But in the motion-bounce illusion, when you see the two objects coincide, you hear a random sound, but you experience the "bouncing" visual motion. In the motion-bounce illusion, your sensory system reconciles an aberrant sound

by making the image that you see more similar to the visual motion that would typically correspond with that sound (a "bouncing" motion).

6. Conclusion

I have argued against various reasons for thinking that crossmodal cases show that at least some of the content of perception is fundamentally multimodal—that is, reasons for thinking that your experience has, say, constitutively audio-visual content (not just a conjunction of an audio content and visual content). In sections two, three, and four, I presented three different reasons for thinking that content of crossmodal experiences is fundamentally multimodal. My claim was that none of these reasons actually entail the conclusion that crossmodal experiences involve fundamentally multimodal content. These reasons do not show that cases like the ventriloquist effect, the McGurk effect, and the motion-bounce illusion must involve fundamentally multimodal content. In section five, I then tried to make some sense of crossmodal cases without making reference such content. This is just a start, but it yields a general two-pronged approach. The first prong is to evoke unimodal features (such as crunchiness and crispness in the Zampini and Spence case). But a unimodal approach is not in itself sufficient. For as O'Callaghan points out, in crossmodal cases, the inputs in two different modalities conflict because they are predicated of a common source or cause (whether it be an individual, object, or event). The second prong is to posit individuals, objects, and events, conceived of in amodal terms (that is, as modality-independent content—content not shared by the senses, but rather content that outstrips the senses). Such an account steers clear of what I was trying to avoid. Making sense of crossmodal cases does not require us to posit multimodal content.[2]

[2] This paper benefited due to comments from Matthew Fulkerson, Bernard Katz, Eric Liu, Mohan Matthen, Barry C.

Bibliography

Bayne, Tim (forthcoming). "Building Block or Unified Fields: How Should We Model the Unity of Consciousness?"

Bayne, Tim (2009). "Perception and the Reach of Phenomenal Content." *The Philosophical Quarterly*, vol. 59, no. 236, 385-404.

Bayne, Tim (2010). *The Unity of Consciousness*. Oxford: Oxford University Press.

Bayne, Tim and Chalmers, David J. (2003). "What is the Unity of Consciousness?" In *The Unity of Consciousness: Binding, Integration, Dissociation*. Ed. by A. Cleeremans. Oxford: Oxford University Press.

Bailey, T. M. and Hahn, U. (2005). "Phoneme Similarity and Confusability." *Journal of Memory and Language*, 52(3), 339-362.

Clark, Austen (2000). *A Theory of Sentience*. Oxford: Oxford University Press.

Hahn, U. and Bailey, T. M. (2005). "What Makes Words Sound Similar?" *Cognition*, 97(3), 227-267.

Howard, I. P. and Templeton, W. B. (1966). *Human spatial orientation.* New York: Wiley.

Matthen, Mohan (2005). *Seeing, Doing, Knowing: A Philosophical Theory of Sense Perception*. Oxford: Clarendon Press.

Matthen, Mohan, Byrne, Alex, Macpherson, Fiona, Siegel, Susanna, and Smith, Barry (2011). *Description of Activities for a Partnership Development Grant from the Social Sciences and Humanities Research Council of Canada*. Unpublished Grant Proposal.

McGurk, H. and MacDonald, J. (1976). "Hearing Lips and Seeing Voices." *Nature*, 264, 746-748.

Nudds, Matthew (2001). "Experiencing the Production of Sounds." *European Journal of*

Smith, and Charles Spence.

Philosophy, 9:2, 210-229.

O'Callaghan, C. (forthcoming). "Perception and Multimodality." In Eric Margolis, Richard Samuels, and Stephen Stich (Eds.) *Oxford Handbook to Philosophy and Cognitive Science*, Oxford University Press.

O'Callaghan, C. (2008). "Seeing What You Hear: Crossmodal Illusions and Perception." *Philosophical Issues*, 18: 316–338.

Sekuler, R., Sekuler, A. B., and Lau, R. (1997). "Sound Alters Visual Motion Perception." *Nature*, 385, 308.

Smith, Barry C. (2013). "Taste, Philosophical Perspectives." In Pashler, Harold E. (Ed.) *Encyclopedia of the mind*. Thousand Oaks, Calif: SAGE Publications, Inc.

Tye, Michael (1995). *Ten Problems of Consciousness*. Cambridge: MIT Press.

Zampini, M. and Spence, C. (2004). "The Role of Auditory Cues in Modulating the Perceived Crispness and Staleness of Potato Chips." *Journal of Sensory Studies*, 19, 347–363.