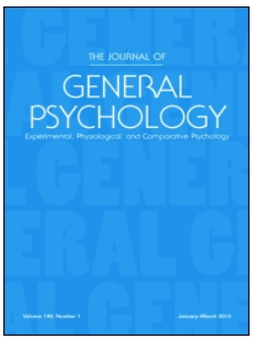

The Journal of General Psychology

ISSN: 0022-1309 (Print) 1940-0888 (Online) Journal homepage: https://www.tandfonline.com/loi/vgen20

McGurk Effect in Gender Identification: Vision Trumps Audition in Voice Judgments

Zehra F. Peynircioğlu, William Brent, Joshua R. Tatz & Jordan Wyatt

To cite this article: Zehra F. Peynircioğlu, William Brent, Joshua R. Tatz & Jordan Wyatt (2017) McGurk Effect in Gender Identification: Vision Trumps Audition in Voice Judgments, The Journal of General Psychology, 144:1, 59-68, DOI: 10.1080/00221309.2016.1258388

To link to this article: https://doi.org/10.1080/00221309.2016.1258388

THE JOURNAL OF GENERAL PSYCHOLOGY
2017, VOL. 144, NO. 1, 59–68
http://dx.doi.org/10.1080/00221309.2016.1258388

Taylor & Francis
Taylor & Francis Group

McGurk Effect in Gender Identification: Vision Trumps Audition in Voice Judgments

Zehra F. Peynircioğlu, William Brent, Joshua R. Tatz, and Jordan Wyatt

American University

ABSTRACT

Demonstrations of non-speech McGurk effects are rare, mostly limited to emotion identification, and sometimes not considered true analogues. We presented videos of males and females singing a single syllable on the same pitch and asked participants to indicate the true range of the voice–soprano, alto, tenor, or bass. For one group of participants, the gender shown on the video matched the gender of the voice heard, and for the other group they were mismatched. Soprano or alto responses were interpreted as "female voice" decisions and tenor or bass responses as "male voice" decisions. Identification of the voice gender was 100% correct in the preceding audio-only condition. However, whereas performance was also 100% correct in the matched video/audio condition, it was only 31% correct in the mismatched video/audio condition. Thus, the visual gender information overrode the voice gender identification, showing a robust non-speech McGurk effect.

ARTICLE HISTORY
Received 6 June 2016
Accepted 3 October 2016

KEYWORDS
Auditory memory; cognition; gender identities; visual

WE INTEGRATE AUDIOVISUAL INFORMATION. If there is a conflict, however, one or the other sense may dominate. Sometimes, audition alters or dominates vision (e.g., Baart & Vroomen, 2010). For instance, our perception of the order in which two lights are presented can be influenced by whether they are presented simultaneously with sounds or are sandwiched between those sounds (Morein-Zamir, Soto-Faraco, & Kingstone, 2003). Other times, vision alters or dominates audition (e.g., Sinnett, Spence, & Soto-Faraco, 2007). For instance, in auditory localization, our perception of where a sound is coming from can be greatly influenced by the spatial location of the corresponding visual information (e.g., Bertelson and Aschersleben, 1998), making such feats as ventriloquism possible.

Perhaps the best-known illusion fueled by the conflict between auditory and visual information is the McGurk effect (McGurk & MacDonald, 1976). The McGurk effect is when auditory perception of linguistic syllables such as /ba/ and /fa/ is altered if one is watching the lips concurrently form another syllable. For instance, if the auditory stimulus is /ba/, but the lips are forming the syllable /ga/,

CONTACT Zehra F. Peynircioğlu peynir@american.edu Department of Psychology, American University, 4400 Massachusetts Ave., Washington, DC 20016-8007, USA.

the auditory perception can become one of /da/. In this case, the auditory and visual information are said to have fused, and the perception is different from either of the presented stimuli. If there is no fusion, then it is more likely for the visual information to influence the percept than the auditory information—participants can report "hearing" the visually presented /ga/ but rarely report hearing the actual auditorily presented /ba/. Similarly, if one has been hearing /ba/ and watching the lips form /ba/, when the auditory stimulus switches to /fa/, but the lips continue forming /ba/, the auditory perception may not switch as it should and one may erroneously still keep "hearing" /ba/. Indeed, demonstrations abound on the Internet, where speech perception is altered by conflicting visual information from the lips. Interestingly, if the question is about the identity of the visual stimulus, then acoustic speech information may alter what participants report seeing whilst lip-reading (Baart & Vroomen, 2010). But, the traditional McGurk effect, as well as our current study, focuses on the identity of the auditory percept and how visual information may alter that percept. A very robust effect, it has been shown across languages (e.g., Magnotti et al., 2015), regardless of whether the syllables are sung or spoken (e.g., Quinto, Thompson, Russo, & Trehub, 2010), and even in infants (e.g., Burnham & Dodd, 2004).

To date, the McGurk effect has been confined largely to the identification of speech sounds or, even if not recognized speech sounds–such as clicks (e.g., Brancazio, Best, & Fowler, 2006)–sounds that can be produced vocally by individuals themselves. One exception is a demonstration by Saldaña and Rosenblum (1993) using plucked and bowed cello sounds and conflicting visual information. However, this effect was considerably smaller than that observed with the identification of speech sounds and present only when the two auditory stimuli were modified for greater similarity. That is, a clear bow sound would not be heard as a clear pluck sound. Other exceptions concern the misperception of emotion when face and voice cues conflict, but the dominance of vision in these cases is less clear, sometimes with the less ambiguous information taking precedence (e.g., Massaro & Egan, 1996), and, though neutral in semantic content, there is still speech present in the auditory cues (e.g., Fagel, 2006). Finally, the dominance of vision over audition under many circumstances is a regularly exploited phenomenon in sound design for film. We do not connect the sounds of film foley artists to their actual sources (e.g., a broom hitting loose leaf newspaper), but rather see them as connected to the content of the film at the corresponding moment (e.g., a bird flapping its wings). But in these cases, the auditory information is somewhat ambiguous and thus not necessarily in conflict with the context set by the visual information. What makes the McGurk effect special and counterintuitive is that misidentification occurs even when the auditory stimulus by itself is not ambiguous and thus when there is an obvious conflict between the two modalities of information presented together.

In the present study, we show a robust non-speech McGurk effect, using voice timbre and visual gender information. We can routinely tell gender information

from just hearing voices (cf. Fellowes, Remez, & Rubin, 1997). To be sure, there are people whose voices surprise us, but even those observations such as "she sounds just like a man" attest to our implicit knowledge and existing limits of expectations from male and female voices. Indeed, auditory speech perception has been shown to be affected by visual or even abstract gender information; for instance, such information can change the heard phoneme boundaries (Johnson, Strand, & D'Imperio, 1999).

Here, we tested whether a voice belonging to one gender (as determined by auditory judgments) would be judged as belonging to the other gender if it emanated from someone of the other gender, thus showing that visual information can override auditory event-identification outside the realm of speech. That is, the judgment would not pertain to the identity of the stimulus content itself as in a typical McGurk study but to the identity of the source of the stimulus. To this end, in order to have a credible way of asking for voice-gender identity without directing the participants' focus onto the now well-known McGurk illusion, we used sung syllables all at the same pitch and asked for an ostensible "natural" voice-range judgment, where a "soprano" or an "alto" judgment would be interpreted as the voice being identified as belonging to a female individual and a "tenor" or a "bass" judgment would be interpreted as the voice being identified as belonging to a male individual. Given that the McGurk effect in speech identification arises regardless of whether the syllables are spoken or sung (e.g., Quinto, et al., 2010), and given that the underlying neural mechanisms in spoken and sung voice perception often recruit the same cortical regions, such as bilateral superior temporal gyri, middle temporal gyri, inferior frontal gyri, and middle frontal gyri (Schön et al., 2010)–albeit somewhat left lateralized in speech and right lateralized in song–this methodology would still enable us to determine whether gender identity based on voice timbre rather than the identity of the utterance itself would also be subjected to a similar illusion.

The theoretical rationale was that if the McGurk effect is indeed limited to speech perception, then gender decisions based on auditory information should not be affected by conflicting visual information as long as the auditory information is held constant and is not ambiguous. However, if, as we hypothesize, the McGurk effect is a manifestation of a more general audio-visual integration phenomenon, then even the perception of the voice timbre that informs one of gender identity, independent of any speech or pitch differences, should be subjected to a similar illusion stemming from visual information overriding auditory information. We should also note that to the extent that the McGurk effect does emerge, however, in this case it cannot be as the fusion of visual and auditory information (e.g., auditorily presented syllable /ba/ and the visually presented syllable /ga/ being fused into a third syllable /da) but as visual information overriding auditory information (e.g., Massaro & Stork, 1998; McGurk & MacDonald, 1976) because our question is on gender identification based on voice quality, and thus a third alternative beyond male or female does not exist.

Method

Participants

Forty-eight American University students participated in the experiment for extra credit in psychology courses. Thirty-two met the inclusion criteria detailed in the Results section.

Materials, design, and procedure

The stimuli were formed using 3- to 4-s video recordings from two mezzo-soprano adult females (vocal range self-reported) and two baritone adult males (vocal range self-reported) singing the syllable /a/ (pronounced "ah") on exactly the same pitch, in this case, a G3 (corresponding to 196 Hz) in a normal fashion for 1 s. The audio was recorded with a Neumann TLM-102 microphone, and digitized at a sampling rate of 44.1 kHz with 16-bit sample resolution. Video was captured with a high-definition camera at 60 frames/s (1280 × 720-pixel resolution). In the video condition, before and after the 1-s duration singing, the singers were shown engaging in body language that signaled they were getting ready to sing, such as taking a preparatory breath, as well as finishing their singing with lip closure. This extraneous visual content, with no sound attached, comprised the non-auditory 2–3 s in each clip. The information in the audio-only condition was taken from the audio portion of the video recordings. Thus, in both the audio-only and video clips, the auditory information was identical.

Each participant first heard two audio recordings (one male and one female) and then saw two video recordings (showing one male and one female). Thus, every participant completed four trials. In the video recordings (created using ProTools), for half of the participants the voice gender in the audio portion of the video matched visual gender actually shown in the video, and for the other half it did not. That is, for the participants in the "mismatch" group, the heard voice gender was different from the visually seen gender. Across four counterbalancing groups in each of the match and mismatch groups, each voice served in the audio and video conditions equally often, and the presentation order was also counterbalanced. No voice was heard more than once by any given participant.

All participants were told that this particular note could be sung by everyone—sopranos, altos, tenors, and basses—but that we wanted to see if the true range of a voice could be identified from its timbre. They were reminded that typically females were sopranos or altos and males tenors or basses, but since crossovers could occur and this note could be sung by anyone, their task was to judge the range of the *voice* in every trial, and regardless of the gender of the person they would see singing in the video conditions. To do so, they were to check one of four boxes corresponding to soprano, alto, tenor, and bass for each of the four voices–in effect making a decision as to whether the voice belonged to a male or a female. Thus, we did not ask for female/male voice decisions but instead inferred gender decisions from the

vocal range decisions. The reason for devising this somewhat indirect measure of gender rather than simply asking for a male/female judgment was to avoid even more obvious demand characteristics or expectations of "getting fooled" by the visual stimulus on the part of participants, which could have influenced judgments even in matched conditions. We should also emphasize that because both groups of participants received the exact same instructions, the exact same audio-only pre-test conditions, and only two video instances, we did not expect any strategy differences or different response biases between the two groups.

Results

We considered the results of the video condition only for those participants who had identified both audio-condition voices correctly by checking either soprano or alto for females and tenor or bass for males and thus giving the correct answer 100% of the time. The reason for asking for perfect performance in the audio condition was to make sure the participant who would see the videos could indeed tell gender from voice information alone. Sixteen participants were excluded from the study because they either did not follow instructions or did not meet the 100% correct performance threshold in the audio condition.

The below results comprise the responses of all of the 32 participants who completed the audio condition successfully. In the matched-video condition, where the gender depicted by the visual information corresponded to that depicted by the auditory information, all 16 participants identified both of the heard voices as belonging to the correct gender. Thus, performance was correct on all 32 of the 32 trials—100% correct. In the mismatched-video condition, where the visual and auditory information conflicted, however, 13 of the 16 participants checked at least one incorrect box (with nine checking the incorrect box in both cases) thus misidentifying the gender depicted by the voice. Thus, performance was correct only in 10 out of the 32 trials—31% correct. Because of the nature of the data, we used the nonparametric Mann-Whitney U test to see whether a difference in the mean rank for number of correct responses existed between the matched- and mismatched-video condition groups. Results showed this difference between the two groups to be highly significant, $U(30) = 232$, $Z = 4.475$, $p < 0.01$, $r = 0.791$ (matched > mismatched). Standard assumptions for this procedure of a single dependent variable, dichotomous independent variables, independence of observations, and heterogeneous distributions were met prior to analysis.

We should also note that the misidentifications in the mismatched-video condition were spread out roughly equally across all four voices and were not limited to "alto" and "tenor." When the visual information depicted a female singing, the first male voice was judged to be a soprano in 12.5% of the trials and an alto in 75% of the trials. The second male voice was judged to be a soprano also in 12.5% of the trials and an alto in 50% of the trials. Similarly, when the visual information depicted a male singing, both female voices were judged to be tenors in 62.5% of the trials and basses in 12.5% of the trials. Finally, as an important side finding, our results

generalized Fellowes, et al.'s (1997) previous findings of successful gender identification based on voice information also to singing voice information.

Discussion

As ubiquitous as the McGurk effect is in misidentifying speech, its demonstrations in the nonverbal arena have been rare (e.g., Saldaña & Rosenblum, 1993) and not without controversy. In this study, we extended the effect to misjudging the identity of the gender of a person's voice rather than to the identity of the content of the spoken speech stimulus. In fact, given that the McGurk effect with speech stimuli does not abate even when the speaker's voice does not match the speaker's picture in terms of gender, at least when the speakers are not familiar to the participants (e.g., Green, Kuhl, Melzoff, & Stevens, 1991), it was interesting to see that when speech identity was kept constant, the illusion could be switched over to the identity of the voice gender itself. Perhaps purists would disagree with calling our findings a McGurk effect because they do not involve speech; however, we believe these findings do satisfy the now generally accepted conceptualization which concerns the misidentification of the identity of any auditory event as a function of conflicting visual information (e.g., Sams, Manninen, Surakka, Helin, & Kättö, 1998). At the very least, they concern how individuals treat audiovisual integration of conflicting information.

To the extent that audiovisual integration (AV) in speech and other auditory events rely on the same underlying mechanism(s), these findings pose added challenges to explanations geared specifically toward speech perception, such as more traditional motor theories of speech (e.g., Liberman & Mattingly, 1985). Such theories describe the McGurk effect as the result of AV streams converging upon a "phonetic module" to produce a single percept based on sympathetic motor articulations. In our study, motor articulations were irrelevant to the task, but when the auditory and visual streams conveying conflicting gender information merged, the end result was a persistence of visual gender information. We should reiterate at this point that because of the lack of options, the typically observed fusion (e.g., the visually presented /ga/ and the auditorily presented /ba/ producing the percept of /da/) was not possible in the present case because there were only two genders the voices could belong to. Instead, visual information trumped auditory information. Thus, although a motor theory could still explain the McGurk effect in speech, an approach that does not separate processing of vocal-tract productions from processing of other auditory events in general would have the advantage of parsimony.

Interestingly, many neuroimaging investigations of the McGurk effect do point to the existence of an AV integration module, but in the left superior temporal sulcus rather than motor cortical areas (e.g., Nath & Beauchamp, 2012). And, in timbre perception in general, the implicated regions are Heschl's gyrus and the superior temporal gyrus (e.g., Halpern, Zatorre, Bouffard, & Johnson, 2004; Menon et al., 2002). Thus, our finding that shows integration of visual information with *voice timbre*

information has the potential to further extend neural representations of AV integration (e.g., Olasagasti, Bouton, & Giraud, 2015). This finding could also provide a starting point for extending behavioral findings in the McGurk effect such as reinvestigating its predictability through eye movements (e.g., Gurler, Doyle, Walker, Magnoff, & Beauchamp, 2015) in broader contexts. That is, just as those individuals who tend to fixate on the visual information provided by the mouth are more prone to showing the McGurk effect, as in the Gurler et al. study, perhaps those individuals who fixate on the visual figure itself are more prone to showing the type of McGurk effect in non-speech events.

The three-step fuzzy logical model of perception (FLMP) proposed by Massaro (1987) would fare better in explaining the current results, whereby an independent evaluation of auditory and visual component features ("feature evaluation"), followed by a synthesis of information from the two modalities ("feature integration"), would then be related to prior knowledge ("decision"). Indeed, FLMP and its newer versions (e.g., Massaro, Cohen, Campbell, & Rodriguez, 2001) are currently the most popular approaches to explaining AV integration and the McGurk effect, and are also consistent with a Bayesian framework (e.g., Massaro & Stork, 1998).

Interestingly, a Bayesian framework has also been used in explaining AV conflicts in multisensory material perception (e.g., Fujisaki, Goda, Motoyoshi, Komatsu, & Nishida, 2014). For instance, in their study where the task was to make likelihood judgments, Fujisaki et al. found that whereas when one type of material was deemed highly more likely upon *hearing* and another type of material was deemed highly more likely upon *seeing*, when the two modalities were presented simultaneously, the decision about the "material" was in line with the *hearing* judgments. Thus, audition ratings dominated vision ratings when there was a conflict. The authors suggested this might be due to an understanding, at some level, that audition is harder to fake and is thus more reliable than vision in the context of material judgments. If we adopt a similar approach as used in the Fujisaki, et al. study, but assume that vision is relied on more heavily in determining gender, this framework also fits our data on category-assignment judgments quite well. That is, when presented with conflicting audiovisual information about gender rather than material composition, individuals may trust visual information more based on previous experiences in such situations.

Indeed, to the extent that gender is a salient part of a person's identity, this assumption that visual information is trusted more in gender determination than auditory information is corroborated by research on speaker identification, as well. Studies of audiovisual integration have shown that individuals are unable to ignore simultaneously presented time-synchronized face information while identifying famous people from their voices (e.g., Campanella & Belin, 2007; Schweinberger, Kloth, & Robertson, 2011; Schweinberger, Robertson, & Kaufmann, 2007). Further, Latinus, VanRullen, & Taylor (2010) have shown that not only does voice gender categorization take longer than face gender categorization despite same levels of accuracy in general, but an incongruency between the simultaneously presented faces and voices decreases only the voice gender categorization accuracy but not the face

gender categorization accuracy. Thus, face processing dominates over voice processing. A similar process can be used to explain why in our study visual information about a person's gender dominated over the actually heard auditory information and created the illusion of the opposite-gender judgments with respect to voice timbre.

One caveat in this Bayesian approach, however, is the seemingly post-hoc assignment of the precognitive Bayesian ratings related to participants' expectations, which determines whether audition or vision will be "trusted" more. The crucial step for predictability in such a model then would be to obtain data on the asymmetry in individuals' reliance on one or the other modality in different situations independent of the data obtained in AV integration tasks. In the present case, the prediction would be that in case of gender identification visual information will be relied on more heavily than auditory information (cf. Latinus, et al., 2010) and thus lead to a McGurk effect when the two modalities of information conflict.

In sum, the lack of an entirely ideal explanation notwithstanding, in this study we have shown a robust McGurk effect that concerns not the identity of the sounds but the source of the sounds (in terms of gender) based on their timbre. When auditory and visual information leading to this identification conflicted, individuals tended to report "hearing" a timbre that was consistent with that implied by the visual information rather than the actual auditory timbre information.

Author notes

Zehra F. Peynircioğlu is a professor of Psychology at American University. She has a BA in psychology from Stanford University, an MA in psychology from Princeton University, and a PhD in psychology from Rice University. She also has a BA in music from American University. **William Brent** is an associate professor of Audio Technology at American University. He has a BA in composition from Wifrid Laurier University, an MA in composition from Mills College, and a PhD in music from University of California, San Diego. **Joshua R. Tatz** is a graduate student in Behavior, Cognition, and Neuroscience at American University. He has a BS in psychology and a BA in music from Drake University. **Jordan Wyatt** is a graduate student in Behavior, Cognition, and Neuroscience at American University. He has a BA in finance from Lehigh University, an MA in counseling from Fairleigh Dickinson University, and an MA in psychology from American University.

References

Baart, M., & Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters*, *471*(2), 100–3.

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, *5*, 482–9.

Brancazio, L., Best, C. T., & Fowler, C. A. (2006). Visual influences on perception of speech and non-speech vocal tract events. *Language and Speech*, *49*, 21–53.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by pre-linguistic infants: Perception of an Emergent Consonant in the McGurk effect. *Developmental Psychobiology*, *45*, 204–220.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543.

Fagel, S. (2006). *Emotional McGurk effect*. Retrieved from https://pdfs.semanticscholar.org/6de6/ed8cadd0359097aca243f49c80f6c2995d67.pdf

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*, 839–849.

Fujisaki, W., Goda, N., Motoyoshi, I., & Komatsu, H., & Nishida, S., (2014). Audiovisual integration in the human perception of materials. *Journal of Vision*, *14*, 1–20.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*, 524–536.

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*, 1333–41.

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*(9), 1281–92.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Audio-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384.

Latinus, M., VanRullen, R., & Taylor, M. (2010). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neuroscience*, *11*, 36.

Liberman, A. M., & Mattingly, I. G. (1985). The motor-theory of speech perception revised. *Cognition*, *21*, 1–36.

Magnotti, J. F., Basu, M. D., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, *233*, 2581–2586.

Massaro, D. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, *8*(1), 1–17.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the Face. *Psychonomic Bulletin and Review*, *3*(2), 215–221.

Massaro, D. W., & Stork, D. G. (1998). Speech Recognition and Sensory Integration: A 240-source-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, *86*(3), 236–244.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–8.

Menon, V., Levitin, D. J., Smith, B. K., Lembke, A., Krasnow, B. D., Glazer, D., & McAdams, S. (2002). Neural Correlates of Timbre Change in Harmonic Sounds. *NeuroImage*, *17*(4), 1742–54.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, *17*, 154–163.

Nath, A. R. & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, *59*(1), 781–787.

Olasagasti, I., Bouton, S., & Giraud, A. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 61–75.

Quinto, L., Thompson, W. F., Russo, F. A., & Trehub, S. E. (2010). A comparison of the McGurk effect for spoken and sung syllables. *Attention, Perception, & Psychophysics*, *72*, 1450–54.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perceptual Psychophysics*, *54*, 406–416.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*(1-2), 75–87.

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J. L., & Besson, M. (2010). Similar cerebral networks in language, music and song perception. *NeuroImage*, *51*(1), 450–461.

Schweinberger, S. R., Kloth, N., & Robertson, D. M. C. (2011). Hearing facial identities: Brain correlates of face-voice integration in person identification. *Cortex*, *47*(9), 1026–37.

Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *The Quarterly Journal of Experimental Psychology*, *60*(10), 1446–56.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: Revisiting the Colavita effect. *Perception & Psychophysics*, *69*, 673–686.