Psychonomic Bulletin & Review
2010, 17 (3), 323-328
doi:10.3758/PBR.17.3.323

Seeing a singer helps comprehension of the song's lyrics

Alexandra Jesse
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

and

Dominic W. Massaro
University of California, Santa Cruz, California

When listening to speech, we often benefit when also seeing the speaker's face. If this advantage is not domain specific for speech, the recognition of sung lyrics should also benefit from seeing the singer's face. By independently varying the sight and sound of the lyrics, we found a substantial comprehension benefit of seeing a singer. This benefit was robust across participants, lyrics, and repetition of the test materials. This benefit was much larger than the benefit for sung lyrics obtained in previous research, which had not provided the visual information normally present in singing. Given that the comprehension of sung lyrics benefits from seeing the singer, just like speech comprehension benefits from seeing the speaker, both speech and music perception appear to be multisensory processes.

Understanding the lyrics of a song is often challenging, not only in contemporary music, but in most vocal performances. Misheard lyrics are coined *mondegreens*, after the classic slip of the ear when the phrase "And laid him on the green" from the 17th-century Scottish ballad "The Bonnie Earl O' Murray" was misperceived as "And Lady Mondegreen" (Wright, 1954). Mondegreens are phonetically similar to the original lyrics and often include misperceived word boundaries (e.g., "laid hi[m]" heard as "Lady"). Mondegreens occur because singing modifies phonetic and prosodic speech properties. For example, in singing, vowels are often lengthened and consonants shortened (McCrea & Morris, 2005; Scotto di Carlo, 2007; Sundberg, 1982). The intelligibility of singing varies also as a function of several other factors, such as pitch level, phonetic context, and singing technique (Scotto di Carlo, 2007). Vowels, for example, become generally less intelligible when sung at a higher pitch (Benolken & Swanson, 1990; Gregg & Scherer, 2006; Scotto di Carlo & Germain, 1985). Intelligibility is also lowered through masking of the singer by instrumental accompaniment. The degree of masking depends on the relative amplitude of instruments and singer and their overlap in their frequency range (Sundberg, 1982).

In the present study, we examined whether the recognition of lyrics is aided by seeing the singer. The primary reason to predict such audiovisual benefit for singing comes from a large literature in speech perception showing a substantial audiovisual advantage. When viewing a speaker's face, perceivers cannot ignore its visual speech information and benefit from its use (Massaro, 1998; Summerfield & McGrath, 1984). Perceivers use visual speech information, even when the auditory signal is not degraded (McGurk & MacDonald, 1976; Reisberg, McLean, & Goldfield, 1987). Seeing a speaker helps recognition in several ways: It provides information about the production of the actual sounds (Miller & Nicely, 1955; Sumby & Pollack, 1954) and about the prosodic structure of an utterance (Bernstein, Eberhardt, & Demorest, 1989; Dohen, Lœvenbruck, & Hill, 2005), which may aid word segmentation (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Importantly, the audiovisual benefit arises from supplementary and from redundant speech information provided by the two modalities (Jesse & Massaro, 2010; Walden, Prosek, & Worthington, 1974). The availability of multisensory information also leads to a processing advantage in nonspeech domains (see Calvert, Spence, & Stein, 2004, for an overview). If the benefit is indeed domain general, it should also be found for singing. We predicted that seeing a singer should provide both supplementary and redundant visual information resulting in an audiovisual benefit for singing.

Our previous experiments involved the alignment of a computer-animated face with the music and lyrics of a contemporary rock song (Hidalgo-Barnes & Massaro, 2007; Massaro & Jesse, 2009). Although the 7% audiovisual benefit obtained for the comprehension of the sung lyrics was a statistically significant improvement (Hidalgo-Barnes & Massaro, 2007), its size was rather small compared with what is commonly obtained for speech (cf. 38% benefit in Jesse, Vrignaud, Cohen, & Massaro, 2000/2001). As Massaro and Jesse (2009) showed, the ample and atypical temporal distortions imposed on the lyrics by the melody of this particular song significantly attenuated the con-

A. Jesse, alexandra.jesse@mpi.nl

tribution of visible speech. A similarly weak benefit was found for a spoken version of the lyrics that had the same temporal distortions.

Synthetic visual speech animations aided the comprehension of temporally undistorted spoken lyrics: When regularly spoken lyrics were heard and seen, an audiovisual improvement of up to 40% was found (Massaro & Jesse, 2009). The weak audiovisual benefit previously observed for singing was thus probably not due to the quality of the synthetic speaker, but due rather to the visible speech distortions when the face was aligned with the sung lyrics. Thus, it is possible that the benefit of seeing visual singing information may be significantly smaller for understanding sung lyrics than for understanding speech. The putative alteration of visual gestures during singing (e.g., see Sundberg, 1982; Thompson & Russo, 2007) might significantly reduce a positive influence of seeing the vocalist. The present experiment tested this hypothesis by presenting sung lyrics along with visual singing.

These weak benefits could, however, also be an artifact of observing visual speech and not visual singing information. In the previous studies, the speaker had always produced visual speech rather than visual singing, even when it was presented along with the original soundtrack of the song (Massaro, 1998). That is, the provided visual information was speech, not singing. Visible singing differs, however, from visible speaking, because it alters articulatory gestures (for a review, see Massaro & Jesse, 2009). For example, lip and jaw openings vary in singing, primarily as a function of pitch (e.g., Sundberg, 1982; Thompson & Russo, 2007). The singing of lyrics should, therefore, look different from the speaking of lyrics. Visible information specific to singing may thus be required to obtain a benefit similar to that observed in speech perception. The observed weak audiovisual benefit in the previous study may have been an artifact due to presenting visual speech rather than visual singing.

In the present study, we investigated whether a substantial audiovisual benefit could be found for lyrics in an ecologically valid situation in which lyrics were sung by a professional singer rather than shown aligned with visual speech. The presented song was from the genre of musical theater, where music and lyrics are generally well-aligned temporally. We predicted that the comprehension of the lyrics would be aided by seeing the singer relative to when the singer can only be heard and, in addition, that this benefit should be similar to that observed in the speech domain (Jesse et al., 2000/2001). A substantial audiovisual benefit for the recognition of sung lyrics would thus provide evidence that visual information aids comprehension similarly in the domain of speech perception and in the perception of singing.

METHOD

Participants

Twenty-six undergraduate students from the University of California, Santa Cruz, participated in return for course credit. All were native speakers of American English, all reported no hearing or language deficit, and all had normal or corrected-to-normal vision.

Materials

Stimuli were taken from a concert video (Mallet, 2001) showing Sarah Brightman singing “Don’t Cry for Me Argentina,” accompanied by a symphony orchestra. The song is part of the musical “Evita,” with music composed by Andrew Lloyd Webber and lyrics written by Tim Rice. The singer is a professional classical soprano opera singer with a three-octave vocal range. We selected 30 unique phrases in which the singer was most visible. Most video clips showed a frontal close-up of the singer; some showed, however, a side view or the singer from some distance. The average length of a phrase was 6.97 words (ranging from 3 to 11 words), or 163.57 NTSC frames. The videos were presented centered on a computer screen in a 640 × 480 pixel display. Video files were audio–video interleave (AVI) containers encoded with the CRAM/msvc Microsoft Video 1 codec. The same video files were presented on auditory-only trials, with the video covered completely by a black rectangle. The audio sampling rate was 16 kHz. Two levels of signal-to-noise (SNR) ratios (-5 and -9 dB) were created by adding white noise to the audio signal in the auditory-only and in the audiovisual presentation conditions. Noise was added in order to avoid ceiling-level performance in the auditory-only condition and, hence, to enable observation of an audiovisual improvement.

Procedure and Design

Participants were tested individually in a sound-attenuated room. Videos were presented at about 50 cm in front of the participant on a 17-in monitor. Audio was presented over Plantronics Audio 90 headphones at a comfortable listening level (76 dB), held constant for all participants. Assignment of participants to one of the two SNR conditions was random. The procedure for these two participant groups was the same. On each trial, participants both saw and heard the singer in the video (i.e., the audiovisual, or AV, condition), saw the video without sound (i.e., the visual-only, or V, condition), or heard but did not see the video (i.e., the auditory-only, or A, condition). Participants were instructed to listen and to watch the computer screen during each presentation and then to type as many words of the lyrics as they thought they had understood. Each participant saw the typed response and was able to make corrections before submitting his or her final response. Trials were self-paced, no feedback was given, and participants were allowed to take a short break between blocks.

Testing consisted of 180 trials, split into two blocks. Each block consisted of three subblocks. In each subblock, all video clips were shown once, with a third of them presented under each modality condition. Across subblocks, the assignment of video clips to modality conditions was counterbalanced in three lists, following a Latin Square design: A–AV–V, AV–V–A, and V–A–AV. The second block was an exact repetition of the first. Trial presentation order within each subblock was, however, always newly randomized. Assignment of phrases to a modality condition and therefore to a list was counterbalanced across participants.

Analyses

To evaluate performance, we used a script to calculate the number of correctly identified words regardless of position in the phrase for each trial. Obvious spelling errors were hand corrected blind to condition. Only complete word matches counted as correct responses. For example, “identify” as a response for “identifies” was scored as incorrect. No partial points were given. Contractions (e.g., “I’ve”) counted as one phonological word. The empirical logit of correctly identified words on each trial was entered in the analyses.

Statistical analyses used logit mixed-effect modeling (Jaeger, 2008), as it is implemented in the lmer function (lme4 package; Bates & Sarkar, 2007) of the R statistical program (R Development Core Team, 2007). P-values were estimated based on Markov chain Monte Carlo simulations ($n = 10{,}000$), with R’s pvals.fnc function (Baayen, Davidson, & Bates, 2008). For all analyses, the best-fitting model was established through stepwise model comparison, using log-likelihood ratio tests. Modality condition (A, AV, and V) and

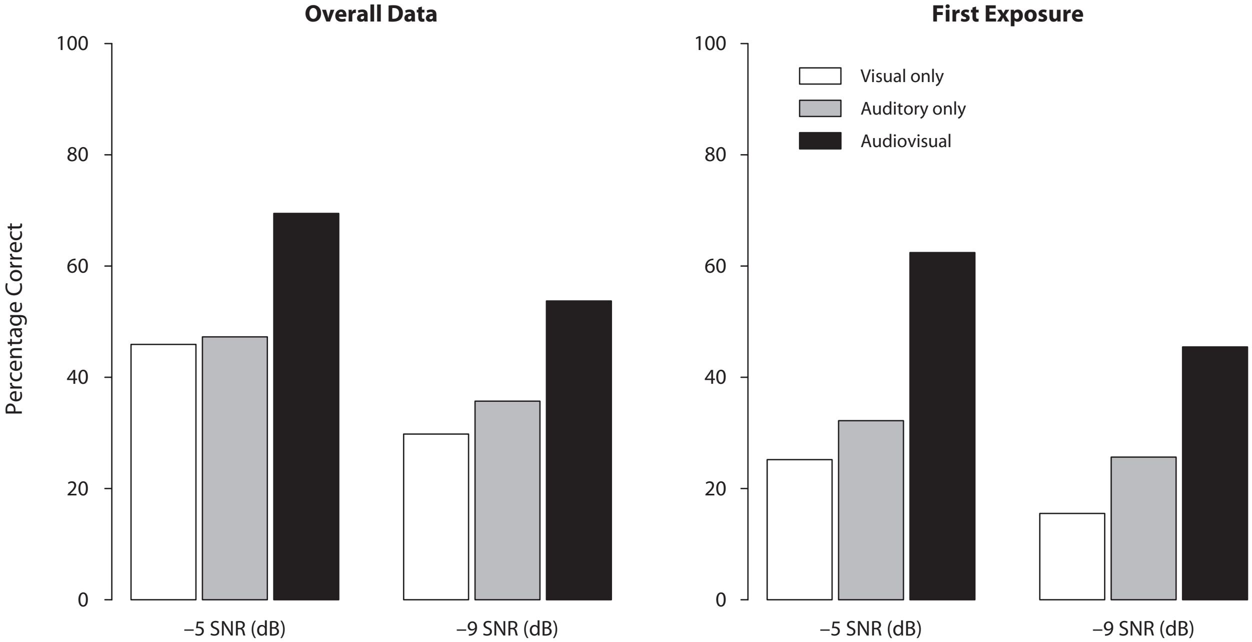

Figure 1. Percentage of correct word responses for each modality condition for each of the two SNRs (-5 and -9 dB) for the overall data set and the first presentation of each phrase only.

SNR (-5, -9 dB) as experimentally manipulated variables and list (A–AV–V, AV–V–A, V–A–AV) as a control variable were treated as categorical fixed factors, for which one condition is mapped onto the intercept. An estimated regression weight indicates the adjustment to be made to the intercept to predict performance at another level of a factor. The sign of the estimate indicates the direction of the adjustment. Repetition was added as a six-step numerical control factor that combined the block and subblock count, centered in its range around 0 (-2.5 to 2.5). Its estimate indicates how the logit of recognized words changes as a function of this predictor. Mixed-effect models allow for the inclusion of participant and item as random effects. All models reported included these two random effects and, therefore, allowed for adjustments to the regression weight estimates on the basis of a participant's or an item's overall mean, with the constraint that the sums of these adjustments must equal zero.

RESULTS

Figure 1 shows the average proportion of correctly identified words as a function of modality and SNR. At both SNRs, more words were correctly recognized when lyrics were presented audiovisually than when they were presented only auditorily (mean percentage of correctly recognized words: -5 dB SNR: $M_{\mathrm{AV}} = 69\%$, $M_{\mathrm{A}} = 47\%$, $M_{\mathrm{V}} = 46\%$; -9 dB SNR: $M_{\mathrm{AV}} = 54\%$, $M_{\mathrm{A}} = 36\%$, $M_{\mathrm{V}} = 30\%$). Phrases were comprehended fairly well when only visual information was provided. The overall audiovisual benefit was 35%, calculated as the mean percentage in the audiovisual condition normalized on performance for auditory-only presentations $(M_{\mathrm{AV}} - M_{\mathrm{A}})/(1 - M_{\mathrm{A}})$, as is commonly calculated in the audiovisual literature (Sumby & Pollack, 1954; but see Ouni, Cohen, Ishak, & Massaro, 2007). Figures 2 and 3 show that the audiovisual benefit was robust across participants and phrases. Although there was some variability across items, this variability was consistent across modalities ($r_{\mathrm{A-AV}} = .83, p < .0001$; $r_{\mathrm{A-V}} = .46, p = .01$; $r_{\mathrm{AV-V}} = .76, p < .0001$).

The best-fitting model consisted of main effects of modality, list, repetition, and SNR and allowed for an interaction between modality and repetition. Lyrics were significantly better recognized when the singer was seen and heard than when she was only heard ($b = 0.9991, p < .0001$, with the auditory-only condition for list A–AV–V, with an SNR of -5 dB mapped onto the intercept with a value of $b = -0.0943$). This benefit did not change over repetitions ($b = -0.0574, p = .09$). Hearing the singer led to better performance than did only seeing the singer ($b = -0.1319, p = .01$). This difference did not vary over repetitions ($b = 0.0507, p = .13$). Overall, however, performance became better with phrase repetitions ($b = 0.1879$, $p < .0001$). Performance was also better with the lower SNR ($b = -0.7086, p = .02$).

The order of modality conditions under which a phrase was presented influenced performance, but did not interact with modality condition. A competitor model allowing this interaction did not provide a better fit to the data. Compared with performance for list A–AV–V, performance was overall better for list AV–V–A ($b = 0.1754, p = .002$) and worse for list V–A–AV ($b = -0.1123, p = .049$). In other words, whenever phrases were presented in the A condition before being presented in the AV condition, they were better recognized. Overall performance was worst for phrases when they were presented V first. Importantly, however, presentation order did not affect the size of the audiovisual benefit.

Even though the audiovisual benefit was not affected by repetitions or lists, we conducted planned comparisons on only the first presentation of a phrase (see Figure 2). The

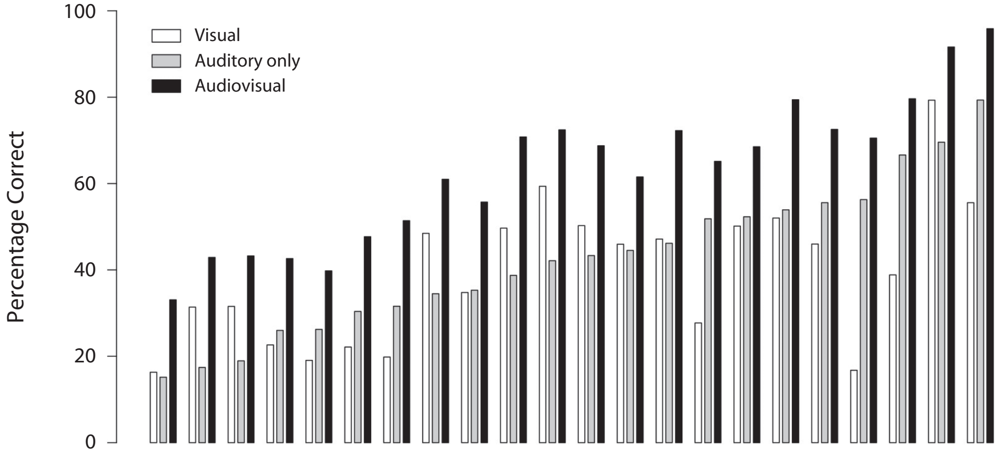

Figure 2. Percentage of correct word responses for each modality condition, grouped by participant and sorted by participants' performance in the auditory-only condition.

normalized audiovisual benefit here was 36%, similar in size to the one found for the overall data set ($M_{\mathrm{AV}} = 55\%$, $M_{\mathrm{A}} = 30\%$, $M_{\mathrm{V}} = 21\%$). The best-fitting model here included only modality as a fixed factor. The analysis thus replicated the audiovisual benefit found for the complete data set ($b = 1.2956$, $p < .0001$, with an intercept estimate of $b = -1.0187$ for the A condition). Performance on auditory-only trials was again better than on visual-only ones ($b = -0.3515$, $p = .006$).

DISCUSSION

The present experiment provides strong evidence that sung lyrics are better understood when the singer is seen as well as heard. This is the first demonstration of this effect when a singer has been shown singing. A substantial audiovisual advantage of approximately 35% recognition improvement was found and was robust across participants and phrases. The first exposure to the song's lyrics already gave a similar-sized audiovisual benefit. Thus, comprehension of lyrics can benefit substantially from seeing the singer, in a manner that is analogous to seeing the speaker in speech perception. The audiovisual benefit is therefore not domain specific to the recognition of spoken language materials but generalizes to sung language materials. The audiovisual benefit in the Hidalgo-Barnes and Massaro (2007) study was small, because a talker was speaking along with, rather than singing along with, the sung lyrics and because of melody-induced temporal distortions in the visual speech. Supporting this, the recognition rate on visual-only presented trials in that study was just 4%. In comparison, we obtained a visual recognition rate of 36% overall and 21% for the first exposure.

With this substantial audiovisual benefit for singing established, future research can focus on the question of what kind of visual information singing provides. Singing should provide some of the same visual information as does speaking. Visual segmental information about consonants, such as their place of articulation, should aid both speech and singing comprehension. Some information available in speaking could be less informative in singing: Lip opening, for example, is a cue to vowel identity in speech, but probably less so in singing, in which lip opening is altered by melody-induced pitch changes (Thompson & Russo, 2007). Nevertheless, our study suggests that the gestures modified by singing generally can still be evaluated in terms of their phonetic content and that they aid in the comprehension of the lyrics.

Some visual information may be more pronounced and may serve a different purpose in singing than in speaking. Vowels are often lengthened in singing (Scotto di Carlo, 2007). Head and eyebrow movements are correlated with changes in pitch in speaking and in singing (Cavé et al., 1996; Munhall et al., 2004; Thompson & Russo, 2007; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). For singing, these facial correlates of pitch may become more important, since they may help in normalizing vowels for melody-induced pitch changes that are larger than those for speech. Pitch-level changes in a sung interval can be estimated by watching a singer (Thompson & Russo, 2007). Whether these facial motion cues affect understanding of lyrics remains to be shown.

The perception of singing involves a language component, but it also contains a musical dimension. The semantic processing of sung lyrics is independent of harmonic processing (Besson, Faïta, Peretz, Bonnel, & Requin, 1998). Phonological processing of vowels, however, is influenced by the processing of the melody of the song (Kolinsky, Lidji, Peretz, Besson, & Morais, 2009). These results are, however, based only on auditory presentations; it remains to be seen to what degree they generalize to the audiovisual processing of singing. Music perception is a multisensory process, as is the perception of speech and, as is shown here, the perception of sung lyrics. The perception of instrumental tone quality is also modulated by visual input. The perception of a tone played by plucking or bowing the cello is influenced by whether one sees plucking or bowing movements, even when perceivers are instructed to ignore what they see (Saldaña & Rosenblum, 1993). Seeing a musician perform provides information about phrasing as well

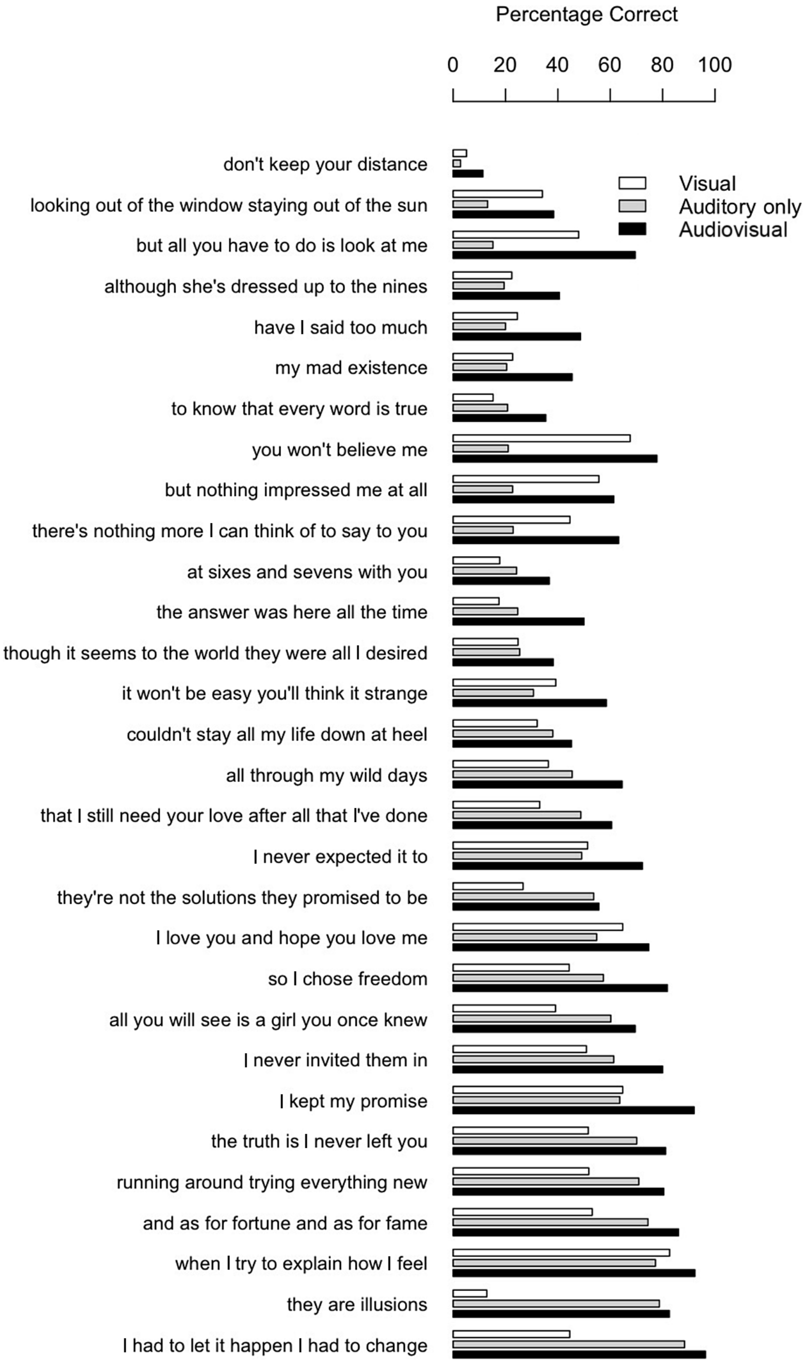

Figure 3. Percentage of correct word responses for each modality condition, grouped by phrase and sorted by performance on the phrases in the auditory-only condition.

as the emotional coloring of music (Dahl & Friberg, 2007; Thompson, Russo, & Quinto, 2008; Vines, Krumhansl, Wanderley, & Levitin, 2006). Perceivers of music performance are also sensitive to audiovisual asynchrony of music stimuli (Vatakis & Spence, 2006; Vines et al., 2006). We can expect that visual processing of speech and nonvocal music share some neurological processes. Watching a pianist play without sound activates in musicians the same brain regions as lipreading does (Calvert et al., 1997; Hasegawa et al., 2004). The present study documents a multisensory nature of perception of musical lyrics analogous to that established for speech perception.

In summary, our results provide the first demonstration that seeing singing improves the understanding of sung lyrics. The present results are striking in that singing alters visual gestures and could therefore have obstructed the recovery of the underlying speech gestures. The audiovisual comprehension benefit is, therefore, not limited to spoken language but extends to sung language and appears to constitute a domain-general phenomenon.

AUTHOR NOTE

We thank Michael Cohen and David Jones for their help with the experiment and Holger Mitterer for discussion. Authors are listed alphabetically because of their equivalent contributions. Please address correspondence to A. Jesse, Max Planck Institute for Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands (e-mail: alexandra.jesse@mpi.nl).

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory & Language*, **59**, 390-412.

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes (R package Version 0.999375-27) [Computer software]. Available at www.r-project.org.

Benolken, M. S., & Swanson, C. E. (1990). The effect of pitch-related changes on the perception of sung vowels. *Journal of the Acoustical Society of America*, **87**, 1781-1785.

Bernstein, L. E., Eberhardt, S. P., & Demorest, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *Journal of the Acoustical Society of America*, **85**, 397-405.

Besson, M., Faïta, F., Peretz, I., Bonnel, A.-M., & Requin, J. (1998). Singing in the brain: Independence of lyrics and tunes. *Psychological Science*, **9**, 494-498.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science*, **276**, 593-596.

Calvert, G. A., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press, Bradford Books.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In H. T. Bunnell and W. Idsardi (Eds.), *Proceedings of the 4th International Conference on Spoken Language Processing* (pp. 2175-2178). Wilmington, DE: Applied Science & Engineering Laboratories.

Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception*, **24**, 433-454.

Dohen, M., Lœvenbruck, H., & Hill, H. (2005). A multi-measurement approach to the identification of the audiovisual facial correlates of contrastive focus in French. In E. Vatikiotis-Bateson, D. Burnham, & S. Fels (Eds.), *Proceedings of the Auditory–Visual Speech Processing International Conference 2005* (pp. 115-116). Adelaide, Australia: Causal Productions.

Gregg, J. W., & Scherer, R. C. (2006). Vowel intelligibility in classical singing. *Journal of Voice*, **20**, 198-210.

Hasegawa, T., Matsuki, K.-I., Ueno, T., Maeda, Y., Matsue, Y., Konishi, Y., & Sadato, N. (2004). Learned audio–visual cross-modal associations in observed piano playing activate the left planum temporale. An fMRI study. *Cognitive Brain Research*, **20**, 510-518. doi:10.1016/j.cogbrainres.2004.04.005

Hidalgo-Barnes, M., & Massaro, D. W. (2007). Read my lips: An animated face helps communicate musical lyrics. *Psychomusicology*, **19**, 3-12.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory & Language*, **59**, 434-446.

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, **72**, 209-225. doi:10.3758/APP.72.1.209

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000/2001). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, **5**, 95-115.

Kolinsky, R., Lidji, P., Peretz, I., Besson, M., & Morais, J. (2009). Processing interactions between phonology and melody: Vowels sing but consonants speak. *Cognition*, **112**, 1-20.

Mallet, D. (Director) (2001). *Sarah Brightman in concert* [DVD]. Los Angeles: Sony Pictures.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press, Bradford Books.

Massaro, D. W., & Jesse, A. (2009). Read my lips: Speech distortions in musical lyrics can be overcome (slightly) by facial information. *Speech Communication*, **51**, 604-621. doi:10.1121/1.1907526

McCrea, C. R., & Morris, R. J. (2005). Comparisons of voice onset time for trained male singers and male nonsingers during speaking and singing. *Journal of Voice*, **19**, 420-430. doi:10.1016/j.jvoice.2004.08.002

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748. doi:10.1038/264746a0

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, **15**, 133-137. doi:10.1111/j.0963-7214.2004.01502010.x

Ouni, S., Cohen, M. M., Ishak, H., & Massaro, D. W. (2007). Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, & Music Processing*, *2007* (Art. No. 47891). doi:10.1155/2007/47891

R Development Core Team (2007). R: A language and environment for statistical computing. (Version 2.6.0) [Computer software]. Vienna: R Foundation for Statistical Computing. Available at www.r-project.org.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). Hillsdale, NJ: Erlbaum.

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

Scotto di Carlo, N. (2007). Effect of multifactorial constraints on intelligibility of opera-singing intelligibility (I). *Journal of Singing*, **63**, 443-455.

Scotto di Carlo, N., & Germain, A. (1985). A perceptual study of the influence of pitch on the intelligibility of sung vowels. *Phonetica*, **42**, 188-197.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio–visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, **36A**, 51-74.

Sundberg, J. (1982). Perception of singing. In D. Deutsch (Ed.), *The psychology of music* (pp. 59-98). New York: Academic Press.

Thompson, W. F., & Russo, F. A. (2007). Facing the music. *Psychological Science*, **18**, 756-757. doi:10.1111/j.1467-9280.2007.01973.x

Thompson, W. F., Russo, F. A., & Quinto, L. (2008). Audio–visual integration of emotional cues in song. *Cognition & Emotion*, **22**, 1457-1470. doi:10.1080/02699930701813974

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters*, **393**, 40-44.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, **101**, 80-113. doi:10.1016/j.cognition.2005.09.003

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech & Hearing Research*, **17**, 270-278.

Wright, S. (1954). The death of Lady Mondegreen. *Harper's Magazine*, **209**, 48-51.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, **30**, 555-568. doi:10.1006/jpho.2002.0165

(Manuscript received June 29, 2009;
revision accepted for publication December 3, 2009.)