Psychomusicology: Music, Mind & Brain
2011, Vol. 21, No. 1 & No. 2

DOI: 10.1037/h0094007

Mouth versus eyes: Gaze fixation during perception of sung interval size

FRANK A. RUSSO
GILLIAN M. SANDSTROM & MICHAEL MAKSIMOWSKI
Ryerson University

ABSTRACT—***We used eye tracking to examine the relative influence of the mouth and eyes on perception of sung interval size. Frequency and duration of gaze were tracked while participants rated the size of intervals produced by two singers in three signal-to-noise conditions, corresponding to high, medium and low audibility. All intervals ascended in pitch direction and ranged in size from 0 to 12 semitones. Both the frequency and duration of gaze fixations revealed that the mouth was the most salient aspect of the visual channel. However, gaze was diverted away from the mouth and toward the eyes with increasing audibility, interval size, and tonal consonance of intervals. A linear regression model incorporating all of these variables accounted for 50% of the variability in gaze duration for the mouth and 45% of the variability in gaze duration for the eyes. Results are discussed in the context of dynamic allocation of attentional resources on the basis of early registration of sensory input. This is the first study of singing to incorporate eye-tracking methodology.***

KEYWORDS—***eye tracking, interval size, cross-modal, visual influences, singing***

Although cognitive science tends to approach music from the perspective of the auditory modality, a number of empirical studies published over the past decade have demonstrated the important role that the visual modality has in shaping our experience of music. Visual aspects of music performance can influence perception of emotion (Dahl & Friberg, 2004; Davidson & Correia, 2002; Thompson, Graham, & Russo, 2005), physiological response (Chapados & Levitin, 2008), perceived tension (Vines, Krumhansl, Wanderley, & Levitin, 2006), and even structural characteristics of music such as perceived note duration (Schutz & Lipscomb, 2007) and sung interval size (Thompson, Russo, & Livingstone, 2010).

The importance of the visual modality in the perception of singing may be owed to a number of factors. These include the natural entwinement of visual and auditory dynamics over the course of song (Thompson et al., 2005), specialized neural circuitry shaped by extensive experience with audio-visual speech (Dick, Solodkin, & Small, 2010), and multimodal mechanisms sub-serving communication that may pre-date song as well as speech (Mithen, 2005, p. 138).

An obvious way in which visual information exerts an influence in perception of singing is in the realm of emotion (Di Carlo, 2004; Thompson et al., 2005). For example, an audio-visual recording of a minor third can be made to convey more happiness if the visual recording is substituted with that of a major third (Thompson, Russo, & Quinto, 2008). The availability of visual information in song is also known to influence perception of phonemes (Quinto, Thompson, Russo, & Trehub, 2010) and comprehension of sung lyrics (Hidalgo-Barnes & Massaro, 2007; Jesse & Massaro, 2010).

One of the more surprising influences of visual information in song is on perception of interval size. An audio-visual recording of a large interval can be made to sound smaller if the visual recording is replaced with that of a smaller interval (Thompson et al., 2010). Remarkably, the effect of

Frank A. Russo, Department of Psychology, Ryerson University; Gillian M. Sandstrom, Department of Psychology, Ryerson University; Michael Maksimowski, Department of Psychology, Ryerson University.

Correspondence concerning this article should be addressed to Frank A. Russo, Department of Psychology, Ryerson University, Toronto, ON M5B 2K3, Canada. E-mail: russo@ryerson.ca

visual information persists even when listeners are (a) asked to focus on auditory information alone and (b) encumbered with a demanding secondary task, suggesting that the visual influence relies upon automatic and pre-attentive mechanisms. One implication of these findings is that gaze behavior responds in a dynamic manner to changes in the availability of auditory and visual information.

Thompson and Russo (2007) investigated the utility of the visual modality for making judgments of interval size by presenting participants with silent videos of singers who sang ascending intervals and by asking participants to rate the size of each interval. The near-perfect correlation observed between rated interval size and veridical interval size implies that observers are able to discriminate intervals on the basis of visual information alone. Motion tracking of the recordings revealed that the maximum displacement of the mouth (separation between lips), eyebrows (raising relative to start position), and head (movement relative to start position) were all positively correlated with sung interval size. Any combination of these visual cues could have thus contributed to the sensitivity to interval size in the absence of auditory information.

Eye tracking is used in the current study to examine the relative influence of the mouth and eyes on judgments of sung interval size. It is possible that both cues are tracked equally or that there is a bias toward one area over the other. Further, the bias may depend on contextual factors. In the case of speech, observers tend to gaze more at the mouth than at the eyes when the task is word identification (Buchan, Paré, & Munhall, 2007) but the pattern is reversed when the task is emotion identification. Because judgments of interval size do not involve emotion in any obvious manner, we predicted that gaze fixations would be biased toward the mouth. Nonetheless, we expected that the eyes would not be entirely neglected because of the important role that emotion has in our everyday experience with song (Welch, 2004, pp. 243-248) and because of the ability of isolated intervals to convey emotion (Cooke, 1959; Thompson et al., 2008).

According to a neuropsychological model of music perception proposed by Peretz and Coltheart (2003), pitch and timing information from the singing voice are fed into an emotion expression analysis module that runs in parallel to a musical lexicon module and a phonological lexicon module. We propose that when an observer is able to see the singer, a default mode is invoked that emphasizes processing in this emotional expression analysis module. Although certain task demands may discourage use of the proposed default mode for audio-visual song (e.g., making a judgment about musical structure), other contextual factors may encourage its use.

Eye-tracking research with speech stimuli has shown that the distribution of gaze fixations can be influenced by the difficulty of the listening conditions. A decrease in audibility due to a reduced signal level or introduction of noise leads to a reduction in the number of transitions between regions of the face (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998) and an increase in the duration of each gaze (Buchan et al., 2007). In accord with the default mode hypothesis and these related findings from speech science, we predicted that the extent of mouth bias would depend on audibility, which was manipulated by varying the extent of noise in the auditory channel during interval presentations.

Participants were asked to make judgments of interval size for audio-visual recordings of sung intervals under three signal-to-noise conditions corresponding to high, medium and low audibility. Gaze fixations were tracked during viewing. If the mouth is prioritized over the eyes in the perception of sung interval size, then there should be more gaze fixations on the mouth relative to the eyes.

METHOD

Participants

16 participants (9 female and 13 right-handed) were recruited from the Ryerson University community. Participants had 0 to 12 years of music training (M = 5.0, SD = 5.1) and ranged in age from 18 to 24 years (M = 19.1, SD = 1.7), and none had received formal vocal training. No participants reported hearing problems, and all had normal or corrected-to-normal vision.

Stimuli and Apparatus

Twenty-six audio-visual recordings were drawn from the stimuli described in Thompson and Russo (2007). The recordings consisted of 13 ascending melodic intervals spanning 0 to 12 semitones produced by each of two female singers (singers 1 and 2 from Thompson & Russo). Both singers were trained pianists with extensive vocal experience in choirs and musical theatre (i.e., 10+ years) but without classical vocal training. Singer 1 had a Russian accent that was mildly perceptible in her singing voice.

Recordings were made using a simple reproduction protocol. Vocalists heard the target interval through headphones. They then attempted to match the pitch and timing of the target, articulating the syllable ''la'' on each note. The mean pitch height of target intervals was centered on the middle of each singer's range, and tones were 1.5 s in duration. The singers were instructed to sing naturally but accurately, and were not informed of the purpose of the experiment. All sung intervals were within 1 cent of the intended interval size. Three signal-to-noise (S:N) versions of each recording were created by first adjusting the original signal to a standard level of 60 dB and then adding white noise to achieve S:N of +8, +4, and 0 dB[1].

Participants were tested in an IAC double-walled sound attenuation chamber. A Tobii x50 eye tracker was used to track gaze fixations. Stimuli were presented using the Tobii visual display and a pair of Logitech loudspeakers (LS11) placed on either side of the display. Participants were seated approximately 30 cm from the display with the centre of the display roughly at eye level. ClearView 2.7.1 software was used to record gaze fixations occurring on each trial, and a custom Matlab script was created to extract and organize the data for analyses. Three regions of interest were defined for each trial (see Figure 1). The eyes region was defined by two rectangles that framed the eyes. The mouth region was defined by a single rectangle that framed the mouth, and the face region was defined by a single rectangle that framed the entire face. Although the regions of interest were static over the course of the trial, margins were large enough to accommodate for head motion. A gaze fixation was defined as a gaze stabilized on any screen location that lasted for a minimum of 100 ms.

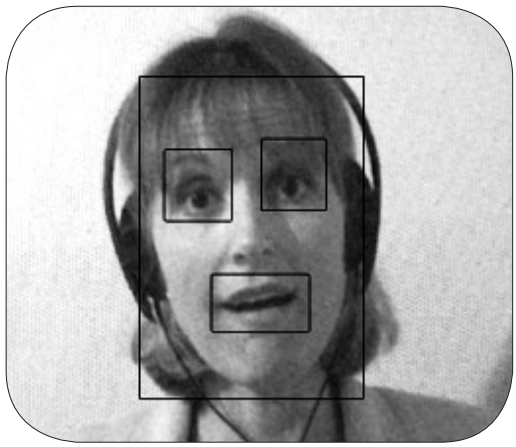

Figure 1. Three regions of interest were specified for each trial. The eyes region was defined by two rectangles. The mouth region and the face region were each defined by single rectangles.

Procedure

Each trial commenced with a 500-ms fixation cross. The fixation cross was followed by a blank screen for 50 ms, which was in turn followed by the audio-visual recording of the interval. The duration of each recording was approximately 4000 ms, including about 500 ms before the onset of the first note and about 500 ms after the offset of the second note. Trials were presented in one of four pre-determined randomized orders, and order was counterbalanced across participants.

We adopted procedures for interval size judgment developed by Russo and Thompson (2005) to encourage responses based on perceptual experience and to discourage listeners with higher levels of musical training from referring to categorical knowledge. First, participants were urged to make their judgments as quickly as possible without compromising accuracy. Second, we utilized a 7-point Likert-type rating scale that made it challenging to easily map the range of chromatic intervals tested (0 to 12 semitones). Third, we

explicitly requested that participants refrain from attempting to map categorical labels onto the Likert-type scale. Finally, we told participants that we would be tracking their response times.

RESULTS

Interval Size Judgments

As seen in Figure 2, rated interval size was significantly correlated with veridical interval size at all levels of S:N, $r(12) = .99$, .99 and .97, ps < .001 (+8, +4 and 0 dB, respectively). The correlation between rated interval size and veridical interval size was computed for each participant at each level of S:N. The strengths of these correlations were subjected to a repeated measures analysis of covariance with S:N and singer entered as the within-subjects factors and with music training entered as the covariate. The effect of S:N was not significant, $F(2, 28) < 1$, and neither was the effect of singer nor the covariate of music training, Fs (1, 14) < 1.

Eye Tracking

Figure 3 shows that the majority of facial fixations fell outside of the mouth and eyes regions, and that a higher proportion of the facial fixations fell on the mouth than on the eyes, $t(31) = 2.64$, $p < .05$. This finding supports the hypothesis that the mouth is prioritized over the eyes in perception of sung interval size. Separate repeated measures analyses of variance (ANOVAs) were carried out for total duration of gaze on the mouth and the eyes, with singer, S:N, and interval as the within-subject variables.

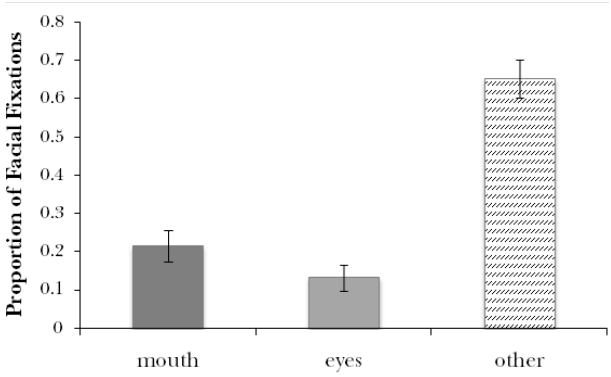

Figure 3. Proportion of gaze fixations falling on mouth, eyes, and other regions of the face.

Mouth. Figure 4 shows that gaze durations increased with decreasing S:N, $F(2, 30) = 3.73$, $p < .05$. Planned repeated contrasts on S:N revealed that gaze durations did not differ between +8 dB and +4 dB, $F(1, 15) < 1$, and that +4 dB led to marginally shorter gaze durations than 0 dB, $F(1, 15) = 3.79$, $p = .07$. Figure 5 shows that gaze durations were longer for singer 1 ($M = 799.01$; $SE = 145.31$) than for singer 2 ($M = 514.22$; $SE = 100.93$), $F(1, 15) = 8.09$, $p < .05$, and that gaze duration varied across interval, $F(12, 180) = 4.21$,

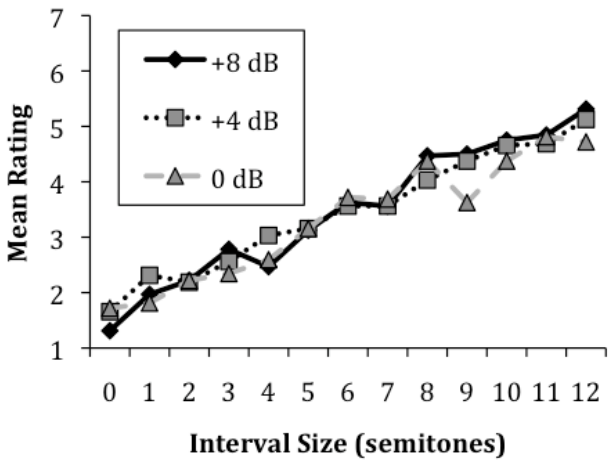

Figure 2. Mean ratings of interval size (on a 7-point Likert-type scale) as a function of veridical interval size across S:N.

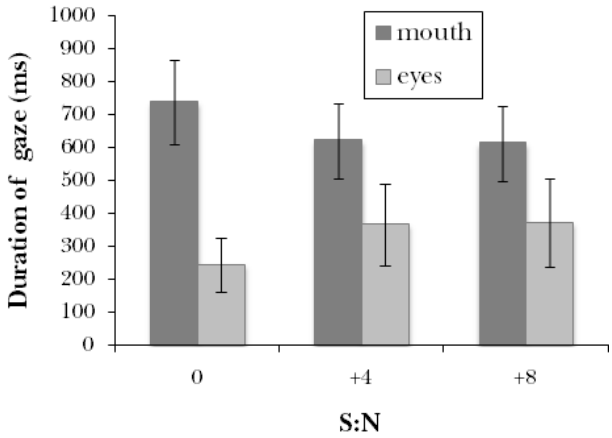

Figure 4. Average duration of gaze on mouth vs. eyes across S:N.

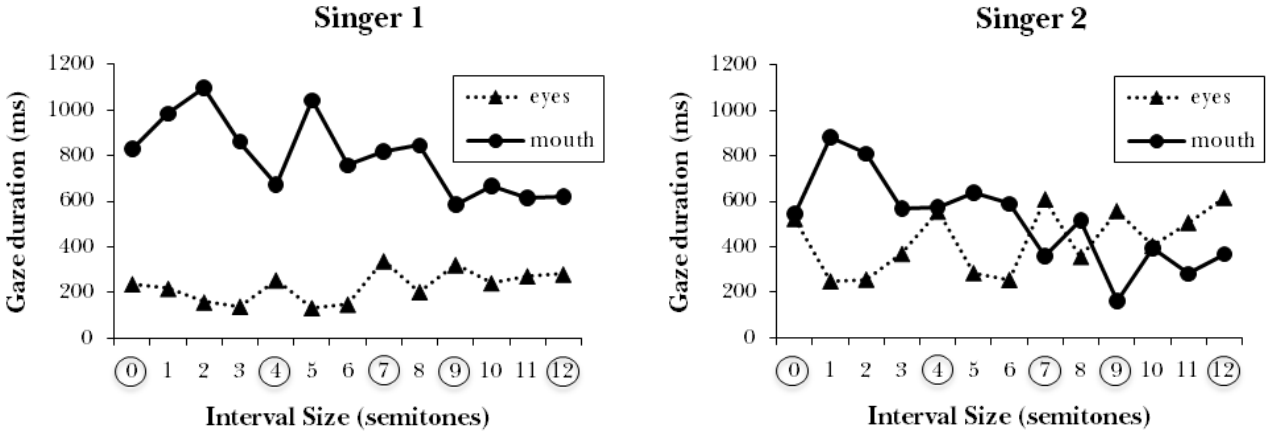

Figure 5. Average duration of gaze on mouth vs. eyes across interval size. Results are plotted separately for singer 1 (panel A) and singer 2 (panel B). The five most tonally consonant intervals are circled.

$p < .001$, with durations generally increasing with decreasing interval size. Although the linear trend for interval was significant, F (1, 15) = 11.17, $p < .01$, the relation between mouth gaze and interval size was nonmonotonic. In particular, positive deviations from the linear trend were observed for the unison, major 3rd, perfect 5th, major 6th and octave (see Figure 5). None of the interactions between terms were significant.

To further explore the effect of interval, we regressed each predictor (singer, S:N, and interval size) onto duration of mouth gaze. The 3-factor model was significant, F (3, 74) = 22.12, $p < .0001$, accounting for 47% of the variation in duration. We next added a fourth factor to represent variation in tonal consonance of the intervals. This predictor was implemented using the formula suggested by Hutchinson & Knopoff (1978), which draws heavily on Plomp and Levelt's (1965) psychoacoustic reworking of Helmholtz's earlier theory. The four-factor model was significant, F (4, 73) = 18.33, $p < .0001$, accounting for 50% of the variation in duration of mouth gaze, with all predictors contributing significantly (see Table 1). The model indicates that duration of mouth gaze increases with decreasing S:N (i.e., as the listening conditions worsen), decreasing interval size and decreasing tonal consonance.

Eyes. Patterns of means were generally the inverse of what was found in the analysis of mouth data. Gaze durations were longer for singer 2 (M = 426.47; SE = 141.02) than for singer 1 (M = 228.03; SE = 88.80), F (1, 15) = 10.12, $p < .01$, and gaze duration increased with increasing S:N (i.e., as the listening conditions improved), F (2, 30) = 5.17, $p < .05$ (see Figure 5). Planned repeated contrasts on S:N revealed that gaze durations did not differ between +8 and +4 dB, F (1, 15) < 1, and that +4 dB led to significantly longer gaze durations than 0 dB, F (1, 15) = 6.17, $p < .05$. Gaze durations also varied across interval, F (12, 180) = 4.26, $p < .001$, with durations generally increasing with increasing interval size. The linear trend for interval was significant, F (1, 15) = 10.05, $p < .01$. None of the interactions between terms were significant.

We further explored the effect of interval, by regressing each predictor onto the mean duration of gaze. The 3-factor model was significant, F (3, 74) =

Table 1
Summary of Four-Factor Regression Analysis Predicting Duration of Gaze on Mouth

Predictor	*B*	*t*	*p*-value
Singer	-.49	-5.90	.001
S:N	-.39	4.49	.001
Interval size	-.18	-2.04	.05
Tonal consonance	-.18	-2.12	.05

Note. Overall model: R = .71; F (4, 73) = 18.34, $p < .001$.

17.47, $p < .0001$, accounting for 42% of the variation in gaze duration. We then proceeded to the second model, which incorporated the same measure of tonal consonance used in the regression analysis of the mouth data. This 4-factor model was significant, F (4, 73) = 14.65, $p < .0001$, accounting for 45% of the variation in gaze duration, with all predictors contributing significantly (see Table 2). The model indicates that duration of eye gaze increases with increasing S:N, increasing tonal consonance, and increasing interval size.

DISCUSSION

Correlations between rated interval size and veridical interval size were nearly perfect across S:N. The strength of these correlations is higher than what has been observed in audio-alone judgments of interval size (Russo & Thompson, 2005) or visual-alone judgments of interval size (Thompson & Russo, 2007), and thus, are quite likely reflective of multimodal gains in perceptual sensitivity. The comparable strength of correlation across S:N suggests that visual information was used to compensate for challenging listening conditions.

Both the frequency and duration of gaze fixations revealed that the mouth was more salient than the eyes. There are a number of reasons why gaze should be biased toward the mouth in the context of an interval size judgment. First, displacement of the lips is larger than displacement of the eyebrows (Thompson & Russo, 2007). Second, displacement of the lips is better correlated with veridical interval size than is eyebrow raising or head movement (Thompson & Russo, 2007). Third, the mouth is tied in a more obvious manner to the sound source (i.e., the periphery of the vocal tract). Fourth, the mouth being more centrally located than the eyes may provide a more robust measure of head movement.

In the context of speech, previous research has shown that gaze is biased toward the eyes when the task is to judge emotion (Buchan et al., 2007) or intonation patterns (Lansing & McConkie, 1999). As proposed earlier, the default mode of processing in singing is thought to be emotional rather than structural. Therefore, when listening conditions are favorable, observers may revert to the default mode even when the experimental task is clearly structural in nature.

Consistent with this default-mode hypothesis, we observed that duration of eye gaze increased (with concomitant decrease in duration of mouth gaze) with increasing interval size and increasing S:N. Observers may have reallocated resources toward the eyes early in the trial, following early registration of sensory input. Larger intervals tend to involve earlier onset of visible movement acceleration. The visual motion perception system is tuned to movement acceleration, meaning that accelerated movement is more visible than movement at constant velocity (Rosenbaum, 1975). Similarly, higher S:N would have led to increased audibility. Such improvements in perceptual clarity (in either the visual or auditory modality) would have been detected early in the trial, potentially leading to a dynamic shift in the allocation of resources.

The main effect of singer indicates that gaze behavior was dependent on which singer was being observed. In particular, observers gazed longer at singer 1 than at singer 2. Examination of Figure 5 also reveals that gaze durations were more evenly distributed between the eyes and mouth for singer 2 than for singer 1. The more distributed pattern of gaze that singer 2 received may once again be a product of perceptual clarity since singer 1 had a non-native accent that was noticeable in her singing

Table 2
Summary of Four-Factor Regression Analysis Predicting Duration of Gaze on Eyes

Predictor	*B*	*t*	*p*-value
Singer	.53	6.06	.001
S:N	.28	-3.19	.01
Interval size	.18	1.96	.05
Tonal consonance	.19	2.01	.05

Note. Overall model: R = .67; F (4, 73) = 14.65, $p < .001$.

voice. The accent may have led to a decrease in audibility, which may in turn have led to an increased bias of gaze on the mouth. Another possibility has to do with differences in expressiveness of the eyes. The maximum displacement of the eyebrows was greater for Singer 2 than for Singer 1.

Across both singers, gaze duration on the eyes and mouth seems to have been influenced by tonal consonance. Intervals that were higher in tonal consonance led to more of a focus on the eyes than did intervals that were lower in tonal consonance. In fact, the overall mouth bias appears to have been reversed for the five most tonally consonant intervals produced by singer 2 (see Figure 5). This tonal consonance effect cannot be accounted for by the movement data reported in Thompson & Russo (2007). It is possible however, that the type of movement was subtle and therefore not captured by displacement measures. In particular, the *gaze of the singer* may have been more directed at the camera (i.e., facing the viewer) for the five most tonally consonant intervals, because all other factors being equal, tonally consonant intervals are easier to sing (Salzer & Schachter, 1969, pp. 4-5; Schön, Lorber, Spacal, & Semenza, 2003). Being the recipient of a direct gaze may be difficult to ignore and may lead to a reflexive shift in gaze toward the singer's eyes (Friesen, Moore, & Kingstone, 2005).

An interesting side-note regarding shifting gaze toward the eyes, has to do with the "bonding neurotransmitter", oxytocin. Experimentally administered oxytocin has been associated with social approach behavior and with increased gaze directed toward the eyes (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007; Gamer, Zurowski, & Büchel, 2010; Guastella, Mitchell, & Dadds, 2008). Although the timescale involved in the current study renders oxytocin an unlikely mediator of eye gaze, it may play an important role in gaze behavior under more naturalistic and extended samples of singing. It has been suggested that oxytocin is regularly released while listening to music as it is in sexual orgasm or other acts that bring ecstatic pleasure and that this may account for the social bonding that music facilitates (Freeman, 1995, as cited in Huron, 2001).

It is important to note that the majority of facial fixations were not directed to the mouth or the eyes (see Figure 3). This point leads to an obvious question: What else were participants looking at? In our review of the fixation data, there were no other obvious candidates (e.g., the nose or the chin), however, the fixations do seem to generally cluster around the middle of the face. Similar findings have been observed in speech studies (Vatikiotis-Bateson et al., 1998), and they highlight a more general problem of eye tracking research. Specifically, points of gaze fixation and visual attention do not necessarily coincide. Because movement can be tracked efficiently in the periphery, these "other" more centrally located facial fixations, may have contributed to interval size judgments via the same displacements captured by motion tracking (i.e., mouth opening, eyebrow raising and head motion).

The results of the current study lead us to suggest that multiple auditory and visual cues exist to support judgments about the relative size of sung intervals. Similar to the way in which diverse cues may be called upon to gauge depth in visual perception (e.g., interposition, binocular disparity), a subset of the available cues in song may be relied upon when the full complement is not available. For example, although the mouth may be the ideal source of information about interval size, gaze is sometimes redirected toward the eyes. When gaze of the observer is directed toward the singer's eyes and away from the mouth such as with tonally consonant intervals, visual input to interval size judgments may rely on eye or even head movement.

Current work in our lab is utilizing point-light displays (e.g., Johansson, 1973) rendered from motion tracking to selectively preserve or eliminate the different sources of visual information. The point light sources are arranged around the eyes and mouth, and participants are required to judge which of two intervals presented in sequence is larger (forced choice). Point-light displays preserve dynamic information while eliminating static visual cues. Preliminary findings show that although performance is superior under full-light conditions, it remains above chance in point-light conditions, suggesting that the dynamics of the visual channel (e.g., movement acceleration) are critical to visual perception of sung interval size.

A limitation and strength of the current study comes form the narrow task prescribed for the participants. By asking listeners to make judgments about the size of isolated intervals, we likely altered the typical course of gaze behavior. However, doing so almost certainly minimized variability, allowing us to observe a high degree of lawfulness in the data.

In sum, the results of this study demonstrate that the mouth provides critical visual input for perception of interval size. Gaze is diverted away from the mouth and towards the eyes with increasing audibility, interval size, and tonal consonance. These findings suggest that the specific attributes of the multimodal signal that are processed depend greatly on contextual factors. Going forward, it seems prudent to consider other contextual factors in visual performance that may influence the allocation of resources. One potential candidate is gross body movement. Do higher levels of head and body movement detract attention from the eyes or do they free up the observer to engage with the eyes in the same manner as do increases in S:N? The two singers tested in the current study had a similar performance style with only a moderate amount of gross body movement. Our results may have varied considerably with the inclusion of singers possessing more embodied styles of performance.

Eye tracking represents a new and powerful methodology for singing research, providing insight about how audiences direct visual gaze. Future work could integrate tracking of singer's gaze along with tracking of audience gaze to better understand the dynamic interplay of gaze that is involved when singers are close enough to make eye contact with their audience. This approach has the potential to greatly advance our current understanding of singing from a cognitive and social perspective and to offer new insights that may eventually inform vocal pedagogy.

[1] The signal level was equivalent across conditions, but the noise level varied. A S:N of 0 dB means that the signal and noise were equal in intensity (i.e., most challenging), whereas +8 dB means that the signal was 8 dB louder than the noise (i.e., least challenging).

REFERENCES

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience, 2*, 1-13.

Chapados, C., & Levitin, D. J. (2008). Cross-modal interactions in the experience of musical performances: physiological correlates. *Cognition, 108*, 639-651.

Cooke, D. (1959). *The language of music.* Oxford: Oxford University Press.

Dahl, S., & Friberg, A. (2004). Expressiveness of musician's body movements in performances on marimba. In A. Camurri & G. Volpe (Eds.), *Gesture based communication in human-computer interaction, Lecture notes in artificial intelligence* (pp. 479-486). Berlin: Springer.

Davidson, J., & Correia, J. S. (2002). Body movement. In R. Parncutt & G. E. McPherson (Eds.), *The science and psychology of music performance* (pp. 237-253). New York: Oxford University Press.

Di Carlo, N. S. (2004). Facial expressions of emotion in speech and singing. *Semiotica, 149*, 37-55.

Dick, A. S., Solodkin, A., & Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and Language, 114*, 101-114.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves "mind-reading" in humans. *Biological Psychiatry, 61*, 731-733.

Freeman, W. J. (1995). *Societies of brains: A study in the neuroscience of love and hate.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Friesen, C. K., Moore, C., & Kingstone, A. (2005). Does gaze direction really trigger a reflexive shift of spatial attention? *Brain and Cognition, 57,* 66-69.

Gamer, M., Zurowski, B., & Büchel, C. (2010). Different amygdala subregions mediate valence related and attentional effects of oxytocin in humans. *Proceedings of the National Academy of Sciences USA, 107*, 9400-9405. doi:10.1073/pnas.1000985107

Guastella, A. J., Mitchell, P. B., & Dadds, M. R. (2008). Oxytocin increases gaze to the eye region of human faces. *Biological Psychiatry, 63*, 3-5.

Hidalgo-Barnes, M., & Massaro, D. (2007). Read my lips: An animated face helps communicate musical lyrics. *Psychomusicology, 19*(2), 3-12.

Huron, D. (2001). Is music an evolutionary adaptation? *Annals of the New York Academy of Sciences, 930*, 43-61.

Hutchinson, W., & Knopoff, L. (1978). The acoustic component of Western consonance. *Interface, 7*, 1-29

Jesse, A., & Massaro, D. W. (2010). Seeing a singer helps comprehension of the song's lyrics. *Psychonomic Bulletin and Review, 17*, 323-328.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201-211.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language & Hearing Research, 42,* 526-539.

Mithen, S. (2005). *The singing Neanderthals: The origins of music, language, mind and body*. London: Weidenfeld and Nicolson.

Peretz, I., & Coltheart, M. (2003). Modularity and music processing. *Nature Neuroscience, 6*, 688-691.

Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America, 38,* 548-556.

Quinto, L., Thompson, W. F., Russo, F. A., & Trehub, S. (2010). The McGurk effect in singing. *Attention, Perception & Psychophysics, 72*, 1450-1454.

Rosenbaum, D. A. (1975). Perception and extrapolation of velocity and acceleration. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 395-403.

Russo, F. A., & Thompson, W. F. (2005). The subjective size of melodic intervals over a two-octave range. *Psychonomic Bulletin & Review, 12*, 1068-1075.

Salzer, F., & Schachter, C. (1969). *Counterpoint in composition: The study of voice leading*. New York: McGraw-Hill Book Co.

Schön, D., Lorber, B., Spacal, M., & Semenza, C. (2003). Singing: A selective deficit in the retrieval of musical intervals. *Annals of the New York Academy of Sciences, 999*, 189-192.

Schutz, M., & Lipscomb, S. (2007) Vision influences perceived note length. *Perception*, *36*, 888-897.

Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica, 156*, 203-227.

Thompson, W. F., & Russo, F. A. (2007). Facing the music. *Psychological Science, 18*, 756-757.

Thompson, W. F., Russo, F. A., & Livingstone, S. (2010). Facial expressions of pitch structure in music performance. *Psychonomic Bulletin & Review, 17*, 317-322.

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

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*, 926-940.

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.

Welch, G. F. (2004). Singing as communication. In D. Miell, R. A. R. MacDonald & D. J. Hargreaves (Eds.), *Musical communication* (pp. 239-259). Oxford: Oxford University Press.

AUTHOR NOTES

Funding for this research was provided by a Natural Science and Engineering Research Council *Discovery* grant awarded to the first author and a Social Science and Humanities Research Council *Major Collaborative Research Initiative* grant: Advancing Interdisciplinary Research in Singing (AIRS SSHRC MCRI) in which the first author is a co-investigator and theme leader. We are indebted to the singers (Anna Volkova and J.C.) for consenting to our experimentation and to Gabe Nespoli and Alex Andrews for technical assistance. We thank Amy Kleynhans and Lisa Liskovoi for assistance with data processing, and Chris Lachine for additional assistance. Finally, we acknowledge the critical feedback of the two anonymous reviewers.

BIOGRAPHIES

Frank Russo

Frank Russo is Associate Professor of Psychology at Ryerson University (in Toronto, Canada), where he directs the Science of Music, Auditory Research and Technology (SMART) lab. Current work in the lab includes vibro-tactile perception of music, cognitively based music information retrieval, the psychology of singing, and vocal-emotional communication. Other notable work includes consultation with U.S. and Canadian Departments of Transportation on locomotive horn effectiveness, and invention of a sensory-substitution technology supporting perception of music by deaf and hard of hearing individuals.

Gillian M. Sandstrom

Gillian M. Sandstrom earned her M.A. in Psychological Science at Ryerson University, studying music cognition in the SMART lab with Dr. Frank Russo. Her M.A. research concerned the effectiveness of music to regulate stress She is currently a PhD student in Social Psychology at the University of British Columbia. Her research interests include the study of weak social ties and their effects on well-being.

Michael Maksimowski

Michael Maksimowski earned his M.A. in Psychological Science at Ryerson University, studying music cognition in the SMART lab with Dr. Frank Russo. His M.A. research concerned cross-modal integration in perception of music. He is currently in his second year of studies as an M.D. student at Jagiellonian University.