tion. He is an NIH grant recipient and brings special expertise in several areas of music perception: rhythm, emotion, performance, and memory.

Dr. Stewart (BA, MSc, Oxford; PhD University of London) is a newly appointed lecturer in Music and Neuroscience at the Department of Psychology, Goldsmiths College, University of London. Within the remit of her position is the establishment of the very first MSc program worldwide in "Music, Mind and Brain." Dr. Stewart carried out postdoctoral work at the Music and Neuroimaging Laboratory at Harvard Medical School as well as within the Auditory Group, Newcastle University Medical School and the Wellcome Department of Imaging Research where she continues to collaborate. Her research has explored the deficits of music perception and cognition and the processing of music notation using neuroimaging (fMRI) techniques.

Volume 19 issue 1 was published in July 2005. Due to unusual circumstances, this issue (Volume 19 issue 2) has been published quite late. For the sake of making *Psychomusicology* current, the present issue is dated Spring 2007. Regardless of its late appearance, I know you will enjoy the quality and diversity of the five articles presented in this issue. "Read My Lips: An Animated Face Helps Communicate Musical Lyrics" (M. Hidalgo-Barnes and D. W. Massaro) investigated the contribution of facial features to the subjects' understanding of lyrics in music. In "The Effect of Familiar Music on the Perception of Other Individuals" F. J. Garivaldis and S. A. Moss explored the effects of background music while their subjects evaluated the resumes of job applicants. As stated in their abstract, "This familiarity might promote trust and thus inhibit a systematic scrutiny of this person [the applicant]." The report by D. George, K. Stickle, F. Rachid, and A. Wapnford "Correlates of Music Preference with Personal Qualities," correlates 30 musical styles with demographics and personal variables of 358 subjects; relating music styles to the positive and negative profiles of their subjects. The fourth article, "The Mood of Rock Music Affects Evaluation of Video Elements Differing in Valence and Dominance" (M. Shevy), is an investigation in the effects of rock music (ominous and happy) "on audience evaluation of two video elements that differed in positive/negative valence and dominance (a positive, dominant character and a negative, secondary world) within a single video." Finally, The intriguing title of the last article (Kai Karma), "Musical Aptitude Definition and Measure Validation: Ecological Validity can Endanger the Construct Validity of Musical Aptitude Tests," certainly should attract the reader's interest!

This issue ends with a book review by A. L. Lehman: *"Component Skills Involved in Sight Reading Music."* The author is Ji In Lee, who in her published dissertation takes on the ominous task of attempting to understand how the several (many?) cognitive dimensions of sight reading function as a unit.

Jack A. Taylor, Editor

Psychomusicology, 19(2), 3-12
©2007 *Psychomusicology*

READ MY LIPS: AN ANIMATED FACE HELPS COMMUNICATE MUSICAL LYRICS

Miguel Hidalgo-Barnes
Dominic W. Massaro
University of California, Santa Cruz

Understanding the lyrics of many songs, not just contemporary songs, is sometimes difficult. Watching the talker's face improves speech understanding when the speech is degraded by noise or by hearing difficulty. To explore whether the face can be similarly helpful in music, 34 phrases from the song "The Pressman" by Primus (1993) were played to thirteen college students. These phrases were aligned with Baldi, a computer-animated talking head. There were three presentation conditions: acoustic presentation of the lyrics, Baldi's mouthing of the lyrics, and the acoustic lyrics aligned with Baldi. For all three conditions, the students were asked to watch and listen and to immediately type the words they thought were being presented. Performance was significantly better in the bimodal condition than in the auditory condition, showing that visual information from the face contributes to the recognition of musical lyrics. Although the contribution of the face was significant, it was somewhat smaller than that found in speech.

A variety of analogies have been drawn between language and music (Besson & Schön, 2003; Bernstein, 1976; Jackendoff, 1992; Lerdahl, & Jackendoff, 1983; Patel, 2003). Both domains might be considered to be forms of communication, although this characterization is somewhat limited in scope. If we define linguistic (literal word aspects of speech) and paralinguistic (nonliteral nonword aspects of speech: pitch, volume, stress, speed) dimensions of communication, we can claim that language emphasizes the linguistic relative to the paralinguistic whereas musical pieces emphasize the paralinguistic relative to the linguistic. In this research, we test whether the visual modality influences music perception in the same way that it does in spoken language perception.

It has been repeatedly shown that many things affect the understandability of verbal communication. In addition to the auditory and contextual information received by listening to a speaker, comprehension is aided visually by being able to see the speaker's face while talking. This phenomenon is most evident in cases where the verbal information is degraded in some way, such as with noise or hearing impairment (Erber, 1972; Kisor, 1990; Massaro, 1987; Summerfield, 1987).

Although both practitioners of speech therapy and speech science were well aware of the potential richness of speech information in the face, the McGurk illusion (hearing inappropriately because of watching the face) captured the imagination of researchers. The McGurk effect or some variant of it has been replicated and studied across different languages (English, Japanese, Dutch, Spanish, French, German, Cantonese, Finnish, and Thai); across eight decades of the lifespan from infancy onward; and from the perception of nonsense syllables to the understanding of prose. Emerging from this impressive body of activity is the robustness of the phenomenon, holding up independently of the intention of the perceiver and also existing in analogous fashion in other domains such as perceiving emotion from the face and the voice (for a review, see Massaro, 1998).

Speech reading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Jordan & Sergeant, 2000; Massaro, 1998, Chapter 14). In addition, people naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about a 1/5 of a second (Massaro, 1998, Chapter 3). These findings indicate that speech reading is highly functional in a variety of nonoptimal situations.

Complementarity of auditory and visual information simply means that one of the sources is most informative in those cases in which the other is weakest. Because of this, a speech distinction can be differentially supported by the two sources of information. That is, two segments that are robustly conveyed in one modality would be relatively ambiguous in the other modality. For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually.

Perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner (Massaro, 1987; Massaro & Stork, 1998). There are many possible ways to treat two sources of information: Use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from bimodal speech to perform as efficiently as possible (Massaro, 1998, Chapter 4).

One might question why perceivers integrate several sources of information when just one of them might be sufficient. Most of us do reasonably well in communicating over the telephone, for example. Part of the answer might be grounded in our ontogeny. Integration might be so natural for adults even when information from just one sense would be sufficient because audiovisual information is often presented together in nature and we have evolved to utilize this natural occurrence, and/or during development, there is much less information from each sense and therefore integration is all the more critical for accurate performance.

Irrespective of traditional arguments for the primacy of vision, the visual element of speech also contributes to the enjoyment of entertainment. Movies would not be nearly as popular if they offered only the audio component. Our society places a high premium on attending events in person. Whether it is a sporting event or a concert, the experience of being there and seeing what is happening with your own eyes is very important to the human experience. When one takes factors such as these into account, the scientific and economic value of accurately recreating all aspects of human communication in art and music becomes apparent.

The value of visible speech, emotion, and intention in face-to-face communication was the primary motivation for the development of Baldi, a 3-D computer animated talking head shown in Figure 1, as viewed in the present experiment. Baldi provides realistic visible speech that is almost as accurate as a natural speaker (Cohen, Walker, and Massaro, 1996; Cohen, Massaro, & Clark, 2002; Massaro, 1998, Chapter 13). The quality and intelligibility of Baldi's visible speech has been repeatedly modified and evaluated to accurately simulate naturally talking humans (Massaro, 1998). Baldi also has teeth, tongue and palate to simulate the inside of the mouth, and the tongue movements have been trained to mimic natural tongue movements (Cohen, Beskow, & Massaro, 1998).

Baldi's visible speech can be appropriately aligned with either synthesized or natural auditory speech. The goal of the present study was to lipsync Baldi to the lyrics of a song, and to test the extent to which perceiving Baldi will help or hinder a listener's comprehension of the song's lyrics.

Method

Participants

Thirteen students from the undergraduate Psychology participant pool at the University of California at Santa Cruz participated in this experiment. All the students reported that they were native English speakers and that they did not have known hearing defects. The participants fulfilled a course requirement. Their ages ranged from 18 to 21 (mean age: 18.85 years). There were 4 males and 9 females.

Stimuli

The auditory component of this experiment consisted of a series of 34 one- to three- second verses from the song "The Pressman" by Primus (1993). The original lyrics were taken from the song "The Pressman" on the audio CD "Pork Soda" (Original Release Date: April 20, 1993; Label: Interscope Records, ASIN: B000001Y5P). Primus is a San Francisco band who blends a variety of genres including punk, metal, funk and prog-rock with Zappa-esque lyrics. The lyrics are sung by a single vocalist Les Claypool, and the accompanying music includes two guitars and a drummer. This song was selected because the music is unknown to most people (Amazon.com Sales Rank: number 5,166 in Music), the lyrics were difficult to understand, and they are not emotionally antagonizing. (Example: "By the light of lamp I sit to type"). The Appendix gives the lyrics of the song that were used. As can be seen, there are very few syntactic and semantic constraints between successive lyrics. This property

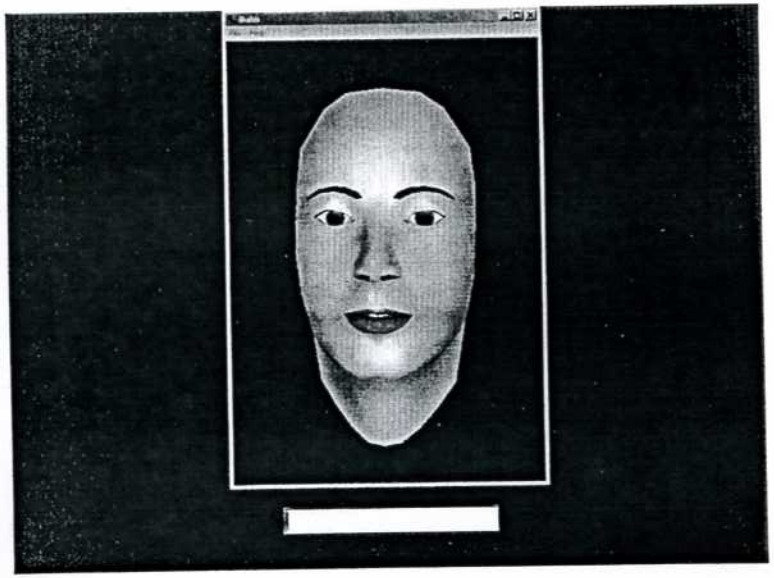

Figure 1. Screen shot of Baldi, as viewed in the experiment. The white box below Baldi shows the words as they are typed in by the participant. On auditory alone trials, only the box was present on the screen.

was desirable because the 34 individual lines or phrases were presented randomly during the experiment. Even if the random presentation of the lyrics made the task more difficult, the experimental hypothesis is testable independently of difficulty level.

These acoustic samples were synchronized with a speech alignment program in the Center for Spoken Language Understanding (CSLU) speech toolkit (http://cslu.cse.ogi.edu/toolkit/). The program takes a text file and corresponding wave file and provides a rough approximation of the location of the phonemes in the sound sample. The alignment was then hand adjusted such that Baldi visually portrayed each phoneme that occurred in each word of the lyrics at the same time and for the same duration as the auditory phonemes present in the song recording. Examples of these test trials can be found at http://mambo.ucsc.edu/psl/primus.html.

At a typical viewing distance of about 18 inches, Baldi's image subtended a visual angle of roughly 15 degrees in height and 7.5 degrees in width. The music and lyrics were presented at a comfortable listening intensity.

Procedure

Participants were tested individually in separate sound isolated rooms. In each of two sessions, the participant was shown each of the 34 samples once in each of the three modalities, Auditory only (A), Visual only (V), and both Auditory and Visual (AV), for a total of 102 trials per session. These 102 unique trials were randomly presented within the session. The two sessions were separated by a 5-minute break. The trials were self paced with each session taking about 30 minutes to complete. In each trial the task of the participant was to listen and watch a computer monitor and then type any words that they understood, and then hit the enter key to go on to the next trial.

Apparatus

The stimuli were presented on PCs running the Windows 2000 operating system with Open-GL video cards, 17 inch video monitors, and sound blaster audio. All the experimental trials were controlled by a CSLU Toolkit RAD application.

Results

Before scoring for accuracy, the typed words were corrected for obvious spelling errors. However, semantically or syntactically confused words (e.g., *bap* for the word *tap*) were scored as incorrect. The proportion of words correctly recognized regardless of position in the lyric was computed for each participant for each of the experimental conditions: session, verse, and presentation condition (A, V, AV). An ANOVA was carried out on the proportion of words correctly recognized (pooled across verse) as the dependent variable and the independent variables of session and presentation condition (A, V, AV). The results show that the presence of the face did indeed help in lyric comprehension. The participants were able to understand 28% and 4% of the lyrics with just the auditory and visual lyrics, respectively, whereas performance was 33% in the bimodal presentation, $F(2, 24) = 527, p < .001$. A specific comparison between the A and AV conditions was statistically significant, $F(1, 12) = 52.9, p < .001$.

Performance also improved from 19% correct in the first session to 24% in the second session, $F(1, 12) = 65, p < .001$. The amount of improvement was somewhat greater for the A condition than for the V and AV conditions, $F(2, 24) = 4.67, p < .02$.

Figure 2 illustrates the individual participant results for the three conditions. As can be seen in the figure, there is a reasonable range of performance across the 13 participants, and some persons benefited more from the presence of the face than others. However, each participant showed an overall advantage of having Baldi aligned with the verses relative to the single modality conditions.

A second ANOVA was carried out on the proportion of words correctly recognized (pooled across sessions) as the dependent variable and with pre-

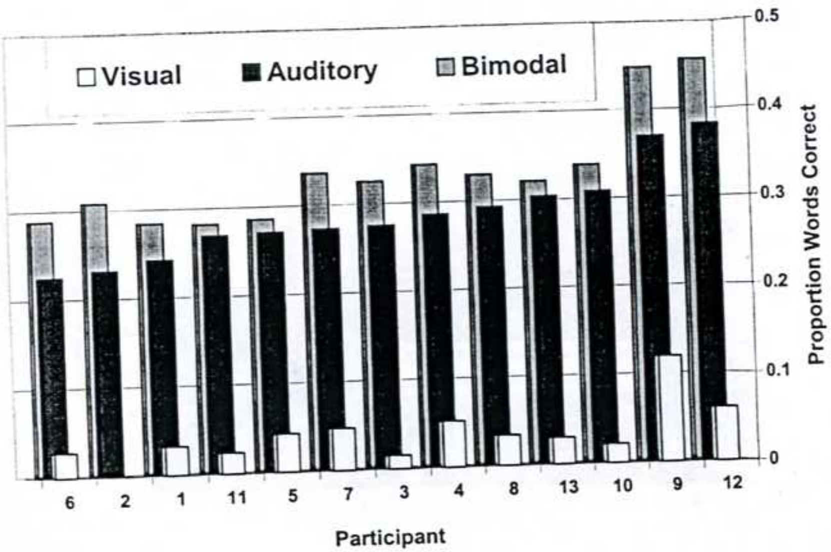

Figure 2. The individual participant results for the three conditions: Auditory only (Auditory), Visual only (Visual), and both Auditory and Visual (Bimodal). The results are arranged from left to right according to performance on the Auditory condition.

sentation condition and verse as the independent variables. There was a large effect of verse, $F(33,396) = 19.7, p < .001$, and an interaction between verse and presentation conditions, $F(66,792) = 13.6, p < .001$. Figure 3 gives the individual verse results for the three conditions. As can be seen in the figure, there is a fairly broad range of performance across the 34 verses, and the face was more effective in some of the verses as compared to others. As can be seen in Figure 3, the accuracy across the verses ranged from 3% ("the little lady said boy you'll never have to be alone") to 58% ("I don't see the sun much these days"). The verse that was repeated (17 and 18) gave fairly similar performance in its two renditions.

Discussion

The research demonstrated that a computer animated face, Baldi, facilitated the recognition and short term memory for a song's lyrics. Although the effect was highly significant, the improvement in performance of 28% to 33% was relatively small. The contribution of the face to the intelligibility of spoken sentences has been found to be much larger. In one study with 71 students (Jesse et al., 2000/01), the test items consisted of 65 meaningful sentences, e.g., "We will eat lunch out." The participants were asked to watch and listen to each sentence embedded in speech noise, and to type in as many words as they could identify. As in the present study with musical lyrics, the sentences were either presented alone or aligned with Baldi. However, a visual alone condition was not included as it was in the current experiment.

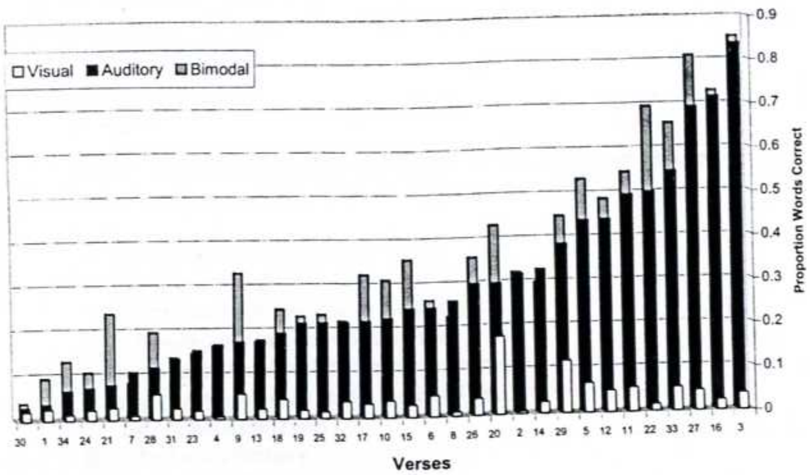

Figure 3. The individual verses results for the three conditions: Auditory only Auditory), Visual only (Visual), and both Auditory and Visual (Bimodal). The results are arranged from left to right according to performance on the Auditory condition. The numerical labeling of the lyrics corresponds to those listed in the Appendix.

Figure 4 shows the performance accuracy in these two conditions for each of 71 participants. As can be seen in Figure 4, the proportion of correct words was higher for the bimodal than the unimodal condition for all 71 participants. If the auditory speech was aligned with Baldi, the participants recognized an average of 66% of the words. Without this additional visual information, recognition was only 45%. When we computed the advantage given by Baldi across different levels of performance, there was an advantage of about 34% for participants at about the same level of performance as those in our current study.

Future studies will have to address why the sentences in noise benefited so much more by the presence of Baldi than did the musical lyrics. Unfortunately, a visual alone condition was not included in the speech in noise study as it was in the current music experiment. It remains possible that the visual speech was much more informative than the visual lyrics and therefore, the

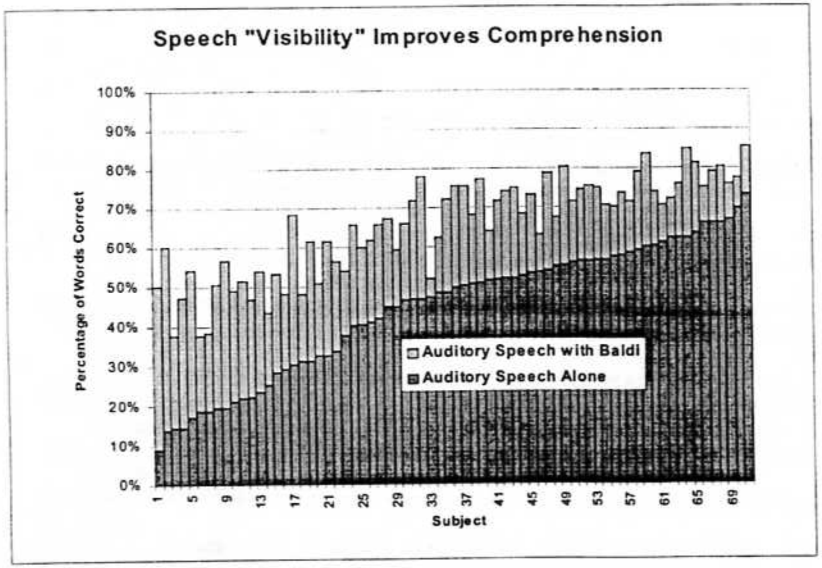

Figure 4. Proportion of words correctly reported for auditory speech alone and auditory speech plus Baldi conditions (from Jesse et al., 2000/2001).

bimodal advantage relative to the auditory condition would necessarily be larger in the speech than in the music studies.

Another issue is how much a real face synchronized with the musical lyrics would have improved performance. We expect that a real face would have given a larger improvement, as is the case in spoken language. In the same study with sentences in noise (Jesse et al., 2000/01), a real face gave about a 46% improvement relative to Baldi's 34% improvement.

The song in this study was chosen for the difficulty inherent in understanding the lyrics. The results garnered might not generalize to other forms of music or even other songs: All the lines were from the same song by the same group. We might also expect that overall performance was affected by the participant's familiarity with rock and roll music. Previous experience with listening and trying to understand rock music might have predisposed the participants to do well at this task. We did not ask the participants how much they listen to rock music, but one can certainly make some assumptions based on the demographic they represent. Only one of the participants reported having heard this particular song, and his results were not outside of the range of the other participants. The learning effect present between session one and two shows that familiarity plays a role in performing this task. More generally, it would be interesting to determine how much familiarity with the musical genre influences overall performance and the advantage of having the face available.

Appendix

The 34 Verses used in the Experiment. Note that 17 and 18 were actually different verses even though they had the same words.

1. by the light of lamp I sit to type
2. my notes on tap at my side
3. I don't see the sun much these days
4. a fluorescent tan covers my hide
5. how much impact shall I have this time
6. my goal today is to reach the deadline
7. I write between lines
8. I deal with fantasy
9. I report the facts
10. give them to me please
11. ham and egg salad on white bread
12. keeps me company on nights like this
13. a pack of mentholated cigarettes
14. keep my air nice and thick
15. when I write words flow like coins from a candy box
16. get out of my way I've got something to say
17. the pulse is beating louder now
18. the pulse is beating louder now
19. the cramps in my hand grow more intense with each tic tic
20. tap tap tap tap tap on the key
21. my social life is at an end
22. so it seems to be
23. why don't I trample on your lawn today
24. I'll take the sky of blue turn over old skies of grey
25. I write between the lines
26. I deal with fantasy
27. I am the pressman
28. acknowledge me
29. mother always told me never stray too far from home
30. the little lady said Boy you'll never have to be alone
31. because you build with fountain pen
32. you create the memory stain
33. you are the press man
34. stand straight boy

References

Besson, M., & Schön, D. (2003) Comparison between language and music. In I. Peretz and R. Zatorre (Eds.), *The cognitive neuroscience of music* (pp.269-293). Oxford: Oxford University Press.

Bernstein, L. (1976). *The unanswered question - six talks at Harvard.* Cambridge. MA: Harvard University Press.

Cohen, M. M., Beskow, J., & Massaro, D. W. (1998). Recent developments in facial animation: An inside view. *Proceedings of Auditory Visual Speech Perception '98*, 201-206. Terrigal-Sydney Australia, December, 1998.

Cohen, M. M., Massaro, D. W., & Clark R. (2002). Training a talking head. In D.C. Martin (Ed.), *Proceedings of the IEEE Fourth International Conference on Multimodal Interfaces, (ICMI'02)* (pp. 499-510). Pittsburgh, PA.

Cohen, M. M., Walker, R. L., & Massaro, D. W. (1996). Perception of synthetic visual speech. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines*, 153-168. New York: Springer.

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15*, 423-422.

Jackendoff, R. (1992). *Languages of the mind.* Cambridge, MA: The MIT Press.

Jesse, A., Vrignaud, N., & Massaro, D. W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. *Interpreting, 5*, 95-115.

Jordan, T., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43*, 107-124.

Kisor, H. (1990) *What's that pig outdoors? A memoir of deafness.* New York: Hill and Wang.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music.* Cambridge, MA: MIT Press.

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

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

Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration. *American Scientist, 86*, 236-244.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience* 6(7), 674-681.

Primus (1993). *The Pressman.* From the Album *Pork Soda.* Santa Monica, CA: Interscope Records.

Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In: *Hearing by eye: The psychology of lip-reading* (Dodd B, Campbell R, Eds). Hillsdale, NJ: Lawrence Erlbaum Associates.

Author Note

This work was supported in part by grants from the National Science Foundation (NSF CHALLENGE Grant CDA-9726363 and NSF Grant BCS-9905176), a grant from the Public Health Service (Grant PHS R01 DC00236), cooperative grants from the Intel Corporation and the University of California Digital Media Program (D97-04), and grants from the University of California, Santa Cruz. The authors thank Ari Patel and the anonymous reviewers of an earlier version of this paper for helpful comments.

Send correspondence to Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064, USA. Electronic mail may be sent to massaro@fuzzy.ucsc.edu.

P:
©

THE EFFECT OF FAMILIAR MUSIC ON THE PERCEPTION OF OTHER INDIVIDUALS

Filia J. Garivaldis
Simon A. Moss
Monash University, Victoria, Australia

When individuals hear recognizable music, they are likely to perceive another person in the environment as more familiar–through a process of assimilation. This familiarity might promote trust and thus inhibit a systematic scrutiny of this person. To assess the effect of recognizable music on the perception and evaluation of other individuals, 87 participants received two hypothetical résumés that were submitted by job applicants. They evaluated the personality of these applicants and designated the candidate they believed was preferable. During this procedure, background music was presented, and this music was familiar to only a subset of participants. Participants who regarded the music as familiar were more likely to perceive the candidates as extraverted. These participants were also more likely to prefer the candidate who demonstrated superior expertise but submitted a less attractive résumé, which suggests that familiar music promoted a systematic rather than superficial analysis.

Music can dramatically influence the decisions, judgments, and behavior of individuals (e.g., Campbell. 1996; Hallam, Price, & Katsarou, 2002; Husain, Thompson, & Schellenberg, 2002; Pates, Karageorghis, Fryer, & Maynard, 2003; Sollberger, Reber, & Eckstein, 2003). For example, soothing music can promote compliance to the requests of salespersons (Chebat, Vaillant, & Gelinas-Chebat, 2000). Indeed, in addition to mood, many features of music can impinge upon the cognition and behavior of individuals, including tempo, timbre, and style (see Bruner, 1990). Thus music is ubiquitous in electronic advertisements as well as other customer interfaces, such as retail stores and even websites.

The music in these contexts is often an extract of a previous recording. Occasionally, however, this music is composed specifically for a campaign or organization. In other words, the extent to which the music is familiar can vary considerably. The effect of music familiarity on the information processing of individuals has, however, received scant attention (although see Rainey & Larsen, 2002). This study underscores some of the benefits and drawbacks that accrue when familiar, recognizable music is presented.

Effect of Recognizable Music on the Perceived Familiarity of Other Individuals

Although the relative merits of recognizable rather than novel music have not been explored comprehensively, a plethora of studies has revealed that familiarity of persons, objects, and contexts can affect the information processing of humans. Indeed, an object or person is typically perceived more