Strategies for Auditory Display of Social Media

Sonifying social-media feeds increases accessibility and enhances cultural meaning and enjoyment for users.

By R. Michael Winters, Neel Joshi, Edward Cutrell, & Meredith Ringel Morris

FEATURE AT A GLANCE:
Social media is an overwhelmingly visual medium, and we ask the simple question: How can the data and images of social media posts be transformed into something as meaningful and vivid in the auditory sense? Such a design would be useful for eyes-free browsing and could enhance the existing visual media. Our strategy first uses artificial intelligence systems to transform low-level input data into high-level sociocultural features. These features are then conveyed using a multifactored temporal design that uses speech, sonification, auditory scenes, and music.

KEYWORDS:
interface design, controls and displays, music, images, universal design, aging, children, accommodation, auditory, media, cultural effects, environment, culture

Imagine opening the feed from your favorite social media platform today. While looking at it, you would see items such as text, photographs, infographics, emojis, gifs, and videos. These items would be arranged in two-dimensional frames, passing by in one direction on a mostly silent screen.

This article starts with a simple question: How might this feed be heard? Then, more specifically, how could the layers of meaning and information embedded in this feed be represented through sonic information design? Although intriguing in its own right, such a design could also be useful in situations where someone wishes to have an experience of the feed without using his or her eyes. Similar to listening to podcasts or music, social media could be used while on the go or while doing other manual activities, like cooking or housework. Sounds could also supplement the existing visual browsing experience.

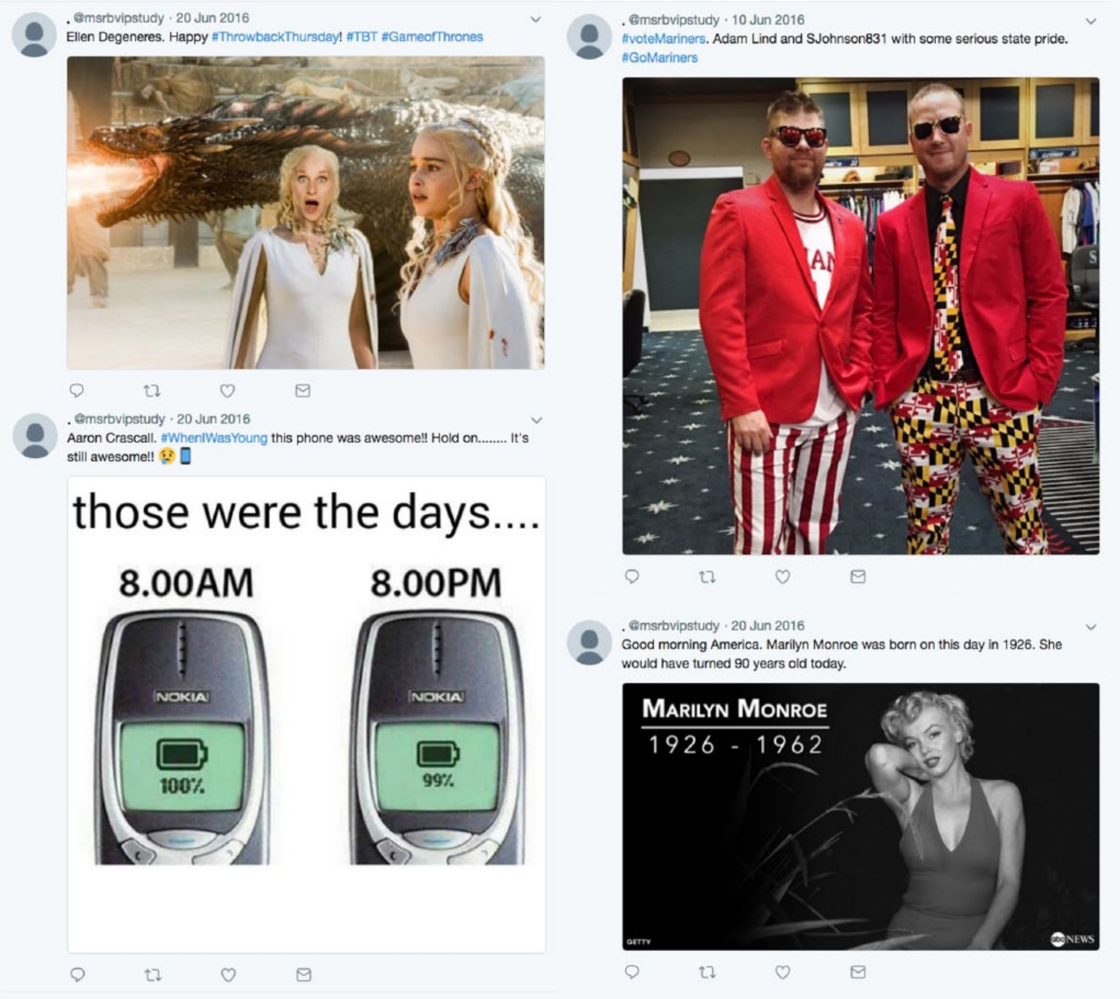
In this article, we present an aspirational design for such a system with two parts. First, we use artificial intelligence systems to transform the raw data of the feed into more meaningful high-level features. Second, we use a heterogeneous mixture of auditory display types to represent this new information content. We believe both strategies will be useful for future work in the area, and we will illustrate our design with examples contrasting current text-to-speech and enhanced versions.

BACKGROUND AND MOTIVATION

Social media is an increasingly important means of civics, communication and learning about the world (Gil de Zúñiga, Jung, & Valenzuela, 2012; Kavanaugh et al., 2012). It is a multimodal medium that draws heavily upon visual culture: representing content and telling stories in many visually compelling ways (Gamson, Croteau, Hoynes, & Sasson, 1992; Lambert & Press, 2013).

In this article, we ask how social media might be transformed into an auditory form – that is to say, represented with sound. When this transformation is objective, systematic, and reproducible, such that it can be used with any feed whatsoever, it is called sonification (Hermann, 2008). Although there are many applications of sonification, including assistive technology, process monitoring, intelligent alarms, data exploration, and aiding movement (Hermann, Hunt, & Neuhoff, 2011), new trends in the field seek to address the aesthetic (Vickers & Hogg, 2006), social (Supper, 2012), and cultural (Barrass, 2012) potentials of the medium. In particular, an open question remains as to how sonifications might become popular and fun for broad audiences (Barrass, 2012) while staying true to its systematic and objective nature (Winters & Weinberg, 2015). Consider the four posts in Figure 1. What would they sound like? We propose that sonification of social media and images provides such an opportunity.

Social media accessibility. Text-to-speech is the standard of Web accessibility (Leuthold, Bargas-Avila, & Opwis, 2008), and many parts of the Web-browsing experience can be made accessible through semantics (Semaan, Tekli, Issa, Tekli, & Chbeir, 2013). However, several challenges have emerged for making the social media experience accessible to visually impaired users, one of which is the increasing pervasiveness of digital images (Morris et al., 2016). With modern machine



. @msrbvipstudy · 20 Jun 2016
Ellen Degeneres. Happy #ThrowbackThursday! #TBT #GameofThrones

. @msrbvipstudy · 20 Jun 2016
Aaron Crascall. #WhenIWasYoung this phone was awesome!! Hold on........ It's still awesome!! 😢📱

. @msrbvipstudy · 10 Jun 2016
#voteMariners. Adam Lind and SJohnson831 with some serious state pride. #GoMariners

. @msrbvipstudy · 20 Jun 2016
Good morning America. Marilyn Monroe was born on this day in 1926. She would have turned 90 years old today.

Figure 1. A collection of four social media posts with images. This article presents an aspirational design for displaying them as audio and includes examples comparing their enhanced versions with the default screen-reader options.

learning systems, it is possible to generate automatic “alt text” or captions for images (Fang et al., 2015), which has been shown to enhance the experience of visually impaired users of Facebook (Wu, Wieland, Farivar, & Schiller, 2017). However, there is still much room for improvement, including conveying appropriate levels of trust in users (MacLeod, Bennett, Morris, & Cutrell, 2017) and using nonspeech audio to create “rich and evocative understandings” of digital images that include a “sense of presence or aesthetic” (Morris, Johnson, Bennett, & Cutrell, 2018).

Approaches to listening to images. There are many ways in which nonspeech audio can be used to understand or represent an image in an eyes-free setting. For example, many sensory substitution devices (SSDs) transform low-level image data, like color, hue, saturation, or brightness, into auditory parameters, like timbre, pitch, or volume (see Hamilton-Fletcher and Ward, 2013, for a review). Audiophotography systems associate images with recorded “sounds of the moment,” which have been shown to enliven photographs and enhance memories (Frohlich & Tallyn, 1999). In the present design, we introduce a fusion of these two types of systems wherein sounds of the moment and music are added retrospectively based upon content automatically detected in the image. Compared with SSDs that operate on low-level image data (e.g., RGB or HSV), we hypothesize that this high-level approach will aid in faster auditory recognition of people, actions, and objects in an image. (Imagine the sonification of pixels in a photograph compared with the sound of a man laughing. Without significant training, a user would recognize the gender and emotion of a person in the image faster in the latter case.)

DESIGN DESCRIPTION

Preprocessing. To explore this design space, we used a collection of tweets that represented a range of topics, including news, politics, celebrities, and humor, all of which contained images. (That collection can be found at https://twitter.com/msrbvipstudy. See MacLeod et al., 2017, for

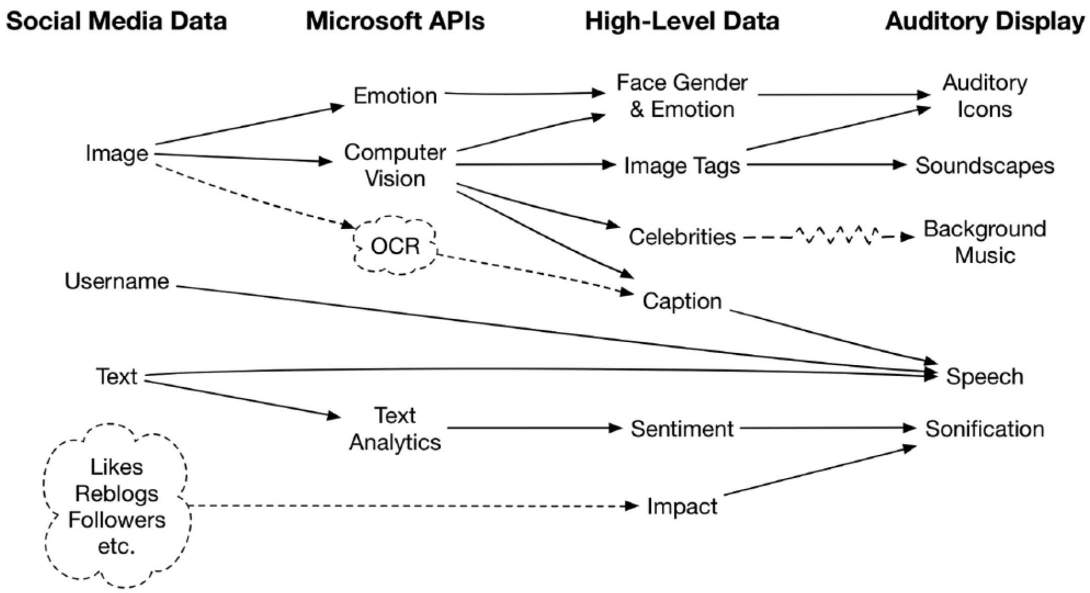


Figure 2. A figure summarizing the transformation. The image and text of the social media post are sent to Microsoft Cognitive Services, which generates socially and culturally meaningful data. These new features become the basis for a multifaceted approach to auditory display incorporating auditory icons, soundscapes, music, speech, and sonification. OCR stands for optical character recognition and is used for reading text on images. The cloud means that the parameter was imagined and not fully implemented at this time.

more details on how this collection was curated.) Data from these posts were extracted and analyzed using several Microsoft Cognitive Service artificial intelligence systems (https://azure.microsoft.com/en-us/services/cognitive-services). In particular, the Computer Vision system was used to generate an automatic description of the image, list objects in the image ("tags"), identify any celebrities, and determine the coordinates and gender of any faces in the image. The Emotion system was used to link the coordinates of each face generated by the Computer Vision system with a value for seven emotion categories. Finally, the Text Analytics system was used to determine the sentiment of the text in each post.

An additional quantity termed *impact* was imagined, designating the relative quantity of activity generated by that post (e.g., number of likes, reblogs). No algorithm or system was available to produce this quantity, so for the purposes of design, its value was equally distributed evenly across the collection. A diagram that summarizes the full high-level transformation is provided in Figure 2.

Mapping strategy. To display the additional information made available by the artificial intelligence systems, the sound design sought to create a coherent balance of the following types of sound:

1. Speech
2. Sonification
3. Auditory scenes
4. Music

Speech was used to speak the text of the user name, main post, and automatically generated image caption. Sonification was used to signal the sentiment and impact using acoustic cues designed to communicate emotion (Winters & Wanderley, 2013). For example, a low-impact positive-sentiment post would sound slow, soft, consonant, and major, and a high-impact negative-sentiment post would sound fast, loud, dissonant and minor. A unique auditory scene – composed of short auditory icons and longer "soundscapes" – was created from the high-level image content. Short, nonspeech auditory icons of males and females expressing various emotions were used to represent the gender and emotion of any faces in the image (e.g., a woman laughing, a man crying), and additional auditory icons represented any sound-producing objects (e.g., a knife sharpening, a bird chirping). Soundscapes were longer in duration and were used to represent recognized actions or environments in the scene (e.g., the sounds of a baseball game or a park). Finally, a random selection of nonvocal background music was used to set the mood of an image, and "theme music" was used if any celebrities were identified. For example, identifying Paul McCartney in the image would trigger playing an excerpt of the song "Band on the Run."

These new auditory components were layered and arranged in time according to three guiding principles. First, we followed a conventional ordering of spoken accessibility content: user name, post text, then image caption. Second, we used layering to minimize the amount of extra time beyond the conventional spoken content. Finally, we began the auditory scenes and music before the main text of the post was spoken – "setting the scene" for the spoken post using the

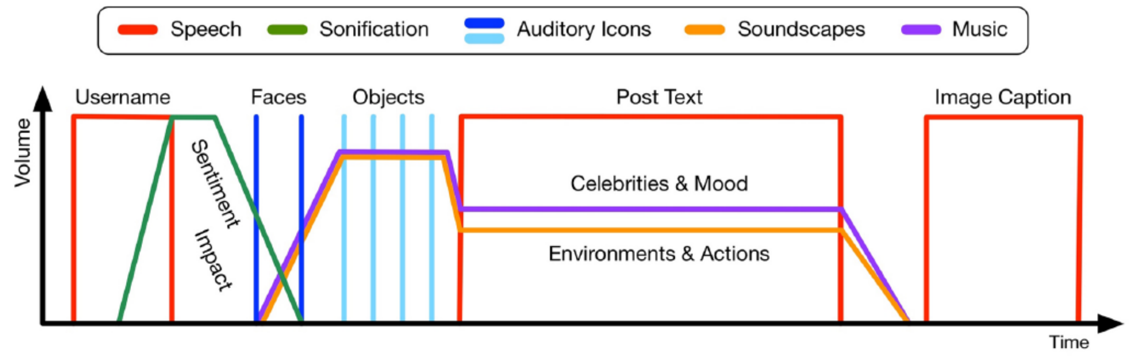


Figure 3. A figure summarizing the approach to the temporal evolution of each post. The user name is spoken first, followed closely by the sonification of sentiment and impact. The sonification fades away as the short auditory icons representing the gender and emotion of faces in the scene are introduced. The music and soundscape fade in, reaching maximum volume as the objects in the scene are introduced. The music and soundscapes then fade into the background as the post text is spoken and finally fade to silence before the spoken image caption.

audio generated by the image. The strategy for temporal evolution reflecting these choices is summarized in Figure 3.

Demonstration. A demonstration of our design applied to a collection of five social media posts is presented in an online video (https://archive.org/details/Auditory_Display_of_Social_Media). The video contrasts text-to-speech versions of the post with the subsequent enhanced version. The first case represents the current speech-only auditory experience, and the second version includes a sonification, auditory scene, music, and image caption.

OPPORTUNITIES FOR AUDITORY CONTENT

We introduce this aspirational design for several reasons. First, we think that social media data provide a clear example where sonification can be applied effectively as a social and cultural medium (Barrass, 2012). Listeners unfamiliar with sonification as a scientific method can still enjoy listening to sonifications of their feed for the objective information it conveys. The layers of information represented in each unique feed, the number of social media users, and the dearth of auditory content in current social media displays make a strong case for its application.

We also think that our data pipeline and mapping strategy point to more general strategies that will be useful in the field. Instead of sonifying low-level image data, like HSV or RGB, we use artificial intelligence as a preprocessing layer to extract high-level image content and enliven the image with automatically generated “sounds of the moment.” In general, we hypothesize that artificial intelligence systems will become an essential component of many auditory display systems, enabling more efficient and cognitively meaningful data to sound mappings.

Finally, our design leverages multiple auditory display types, including speech, sonification, auditory icons, soundscapes, and music. Although the field typically separates these as independent display strategies and does not include music (Hermann et al., 2011), we believe that our mixture of display types is most appropriate for the sociocultural context. All auditory display types have different strengths and weakness, and for data that are heterogeneous in nature, we believe the most compelling display will be created only through thoughtful combination.

CONCLUSION

Social media and images provide a rich context to explore the sociocultural potential of sonic information design, and the future is full of possibilities for the application of auditory display. We hope that our aspirational design and examples can inspire ideas and guide opportunities for this ever-changing information stream.

REFERENCES

Barrass, S. (2012). The aesthetic turn in sonification towards a social and cultural medium. *AI & Society*, *27*, 177–181.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., . . . Zweig, G. (2015). From captions to visual concepts and back. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*. New York, NY: IEEE. doi:10.1109/CVPR.2015.7298754

Frohlich, D., & Tallyn, E. (1999). Audiophotography: Practice and prospects. In *CHI'99 extended abstracts on human factors in computing systems* (pp. 296–297). New York, NY: ACM.

Gamson, W. A., Croteau, D., Hoynes, W., & Sasson, T. (1992). Media images and the social construction of reality. *Annual Review of Sociology*, *18*, 373–393.

Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, *17*, 319–336.

Hamilton-Fletcher, G., & Ward, J. (2013). Representing colour through hearing and touch in sensory substitution devices. *Multisensory Research*, *26*, 503–532.

Hermann, T. (2008). Taxonomy and definitions for sonification and auditory display. In *Proceedings of the 14th International Conference on Auditory Display* (pp. 1–8). Paris, France: IRCAM.

Hermann, T., Hunt, A., & Neuhoff, J. G. (Eds.). (2011). *The sonification handbook*. Berlin, Germany: Logos Verlag.

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., . . . Xie, L. (2012). Social media use by government: From the routine to the critical. *Government Information Quarterly*, *29*, 480–491.

Lambert, J., & Press, D. D. (2013). *Digital storytelling: Capturing lives, creating community* (4th ed.). New York, NY: Routledge.

Leuthold, S., Bargas-Avila, J. A., & Opwis, K. (2008). Beyond web content accessibility guidelines: Design of enhanced text user interfaces for blind internet users. *International Journal of Human–Computer Studies*, *66*, 257–270.

MacLeod, H., Bennett, C. L., Morris, M. R., & Cutrell, E. (2017). Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 5988–5999). New York, NY: ACM.

Morris, M. R., Johnson, J., Bennett, C. L., & Cutrell, E. (2018). Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Paper 59). New York, NY: ACM.

Morris, M. R., Zolyomi, A., Yao, C., Bahram, S., Bigham, J. P., & Kane, S. K. (2016). "With most of it being pictures now, I rarely use it": Understanding Twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5506–5516). New York, NY: ACM.

Semaan, B., Tekli, J., Issa, Y. B., Tekli, G., & Chbeir, R. (2013). Toward enhancing web accessibility for blind users through the semantic web. In *Proceedings of the 2013 International Conference on Signal-Image Technology and Internet-Based Systems* (pp. 247–256). Washington, DC: IEEE Computer Society.

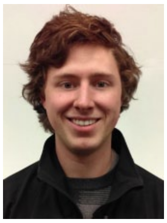
Supper, A. (2012). *Lobbying for the ear: The public fascination with and academic legitimacy of the sonification of scientific data* (PhD thesis). Universitaire Pers Maastricht, Maastricht, Netherlands.

Vickers, P., & Hogg, B. (2006). Sonification abstraite/sonification concrète: An "aesthetic perspective space" for classifying auditory displays in the ars musica domain. *In Proceedings of the 12th International Conference on Auditory Display* (pp. 210–216). Atlanta: Georgia Institute of Technology.

Winters, R. M., & Wanderley, M. M. (2013). Sonification of emotion: Strategies for continuous auditory display of arousal and valence. In *Proceedings of the 3rd International Conference on Music and Emotion* (pp. 492–501). Jyväskylä, Finland: University of Jyväskylä.

Winters, R. M., & Weinberg, G. (2015). Sonification of the Tohoku earthquake: Music, popularization, and the auditory sublime. In *Proceedings of the 21st International Conference on Auditory Display* (pp. 273–280). Graz, Austria: University of Music and Performing Arts Graz.

Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017). Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 1180–1192). New York, NY: ACM.



R. Michael Winters *is a PhD candidate in music technology at the Georgia Institute of Technology, where he works in the Sonification Lab and Robotic Musicianship Group. He has designed sonification systems and mapping strategies for an array of applications, including affective display, smart cities, high energy, thermal and granular physics, medical diagnosis, educational technologies, and social media. His current work studies the neurophysiological and behavioral outcomes of insight during self-referential listening.*

Neel Joshi *is a senior researcher at Microsoft Research. His work is in computer vision, computer graphics, and human–computer interaction, focusing particularly on imaging, computational photography, and creative tools for visual arts. He holds an ScB from Brown University, an MS from Stanford University, and a PhD from University of California–San Diego, all in computer science. He has held internships at Mitsubishi Electric Research Labs, Adobe Systems, and Microsoft Research; was a visiting professor at the University of Washington; and has been at Microsoft Research since 2008.*

Edward Cutrell *is a principal researcher at Microsoft Research, where he explores computing for disability, accessibility, and inclusive design. Over the years, he has worked on a broad range of human–computer interaction topics, including input tech, visual perception and graphics, intelligent notifications and disruptions, and interfaces for search and personal information management. From 2010 to 2016, he managed the Technology for Emerging Markets (TEM) group at MSR India, focusing on technologies and systems useful for people living in underserved rural and urban communities in developing countries. He has worked in the field of human–computer interaction since 2000; he is trained in cognitive neuropsychology, with a PhD from the University of Oregon.*

Meredith Ringel Morris *is a principal researcher at Microsoft Research, where she manages the Ability team. She is also an affiliate professor in the School of Computer Science and the Information School at the University of Washington. More information about her research can be found at http://merrie.info.*

We thank Haley Macleod and Cindy Bennett for their insightful thoughts and contributions as we worked towards the final design. The design described in this manuscript was pursued as part of an internship by the first author at Microsoft Research.

DOI: 10.1177/1064804618788098
Article reuse guidelines: sagepub.com/journals-permissions