How Is Nonverbal Auditory Information Processed? Revisiting Existing Models and Proposing a Preliminary Model

Myounghoon Jeon
Mind Music Machine Lab
Cognitive and Learning Sciences, Computer Science
Michigan Technological University, Houghton, MI

Use of multimodal displays is getting more prevalent in Human Factors and Human-Computer Interaction. Existing information processing models and theories predict the benefits of multimodality in user interfaces. While the models have been refined regarding vision, more granularity is still required regarding audition. The existing models mainly account for verbal processing in terms of representation, encoding, and retrieving, but these models do not provide sufficient explanations for nonverbal processing. In the present paper, I point out research gaps in nonverbal information processing of the representative models at the working memory and attention level. Then, I propose a preliminary conceptual model supported by neural and behavioral level evidence, and provide evaluations of the model and future works.

1 INTRODUCTION

Multitasking is pervasive in our daily lives and we rarely feel such a task is demanding or difficult. Psychological models and theories explain how we can be comfortable with multitasking through our distribution of attentional resources (Wickens, 2002) or processing mechanisms in our working memory (Baddeley, 1992). These models have been better refined in vision, whereas less sophisticated models have been developed in audition despite considerable evidence at neural and behavioral levels, which provides possibilities for a new and improved model of auditory processing. Regarding audition, the models focus more on verbal information processing than nonverbal information processing. Nonverbal auditory information complement, supplement, or even supplant visual information (e.g., ipod Nano). However, use of nonverbal auditory signals in the interface cannot be fully explained by the existing models.

In the present paper, I propose a preliminary conceptual model that can account for nonverbal auditory information processing, including its *representation* and *encoding* processes at the attention and working memory level. This model will contribute to predicting and explaining user behavior in multimodal tasking, specifically, involving nonverbal auditory information. This model will not only guide basic auditory perception researchers but also provide theoretical background to applied multimodal display researchers.

The present paper (1) outlines the existing representative model (working memory model) and theory (multiple resource theory) and highlights their research gaps; (2) addresses the previously proposed alternative model (musical working memory, Berz, 1995) and why it is not sufficient; (3) proposes a new conceptual model and its evaluation; and (4) depicts implications of the model and future works.

2 INFORMATION PROCESSING MODELS

There are many information processing models in cognitive science, including conceptual, qualitative, mathematical, and computational models. Reviewing all models is beyond the scope of the present paper. I selected Baddeley's Working Memory Model (Baddeley, 1992) as a representative model in working memory and Wickens' Multiple Resource Theory (Wickens, 2002) in attention theory. These are among the most widely accepted, taught, and applied the Human Factors community (e.g., in textbooks, Wickens & Hollands, 2000) and can provide explanations for basic and applied research. Furthermore, I will focus only on the part of these models, which is mostly relevant to the discussion of the present paper: auditory information processing.

Working Memory Model (WMM)

The working memory model (WMM) was originally introduced as a three-component system, including central executive, visuo-spatial sketchpad, and phonological loop (Baddeley & Hitch, 1974). Central executive is a main controller capable of attentional focus, storage, and decision making. The visuo-spatial sketchpad and the phonological loop are slave systems. The visuo-spatial sketchpad is assumed to be capable of maintaining and manipulating visual and spatial information. The phonological loop processes auditory information and consists of two components: a phonological store, which holds memory traces in acoustic or phonological form that fade in a few seconds and an articulatory rehearsal process analogous to subvocal speech. The function of the articulatory rehearsal process is to retrieve and re-articulate the contents held in this phonological store and in this way to refresh the memory trace.

The phonological loop is known to process both acoustic and verbal information processing, but most descriptions about representation and encoding focus on

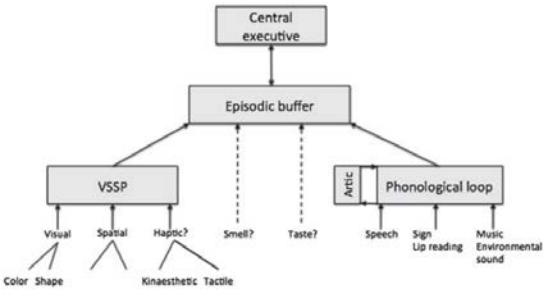

Figure 1. A speculative view of the flow of information from perception to working memory. (Figure 5 in Baddeley, 2012).

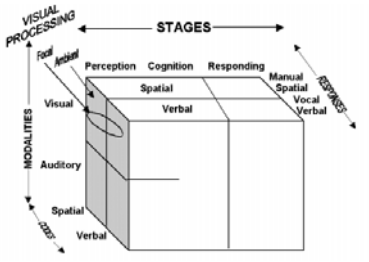

Figure 2. The structure of multiple resources (Figure 8.1 in (Christopher D Wickens & McCarley, 2007).

verbal information processing. In many papers (Baddeley, 1992, 2012) about WMM, the critical research components for the phonological loop have been about language: phonological similarity effect, word length effect, articulatory suppression, and irrelevant sound effects. In contrast, there has been little advance in terms of refining nonverbal auditory information processing. This might be partly because Baddeley himself has actively conducted research on language and also much more researchers have conducted research on speech than non-speech. In more recent research (Williamson, Baddeley, & Hitch, 2006) Baddeley and colleagues suggest that the phonological loop may be capable of holding music sequences. However, this attempt did not refine the model as self-sufficient to account for nonverbal auditory information processing. More specifically, the phonological loop does not detail the "***representation***" format of nonverbal auditory information and how it is rehearsed. Suppose an experiment that can examine representation and encoding of nonverbal information, where people are asked to store several musical notes. If we depend on the phonological loop of WMM, prompt issues we can confront include:

(1) If participants have perfect pitch and are able to recognize the names of each musical note, they might try to rehearse those linguistic labels (e.g., "Do" "Re" "Mi", etc.) for pitch. However, what about other characteristics of the sound? e.g., duration, rhythm, timbre, etc. How are these components encoded in the phonological loop?

(2) What if the sound does not match with musical frequency? (e.g., atonal or micro-tonal music where pitches fall in between the classic 12 step scale). Despite having perfect pitch, participants might not articulate the label.

(3) What if participants do not have perfect pitch or musical knowledge? This may be a more usual case because only 1 in 10,000 people have perfect pitch (Takeuchi & Hulse, 1993). How can the phonological loop help in terms of representing or encoding of the nonverbal information?

(4) According to this working memory model, spatial information is primarily processed in the visuo-spatial sketchpad, but auditory information (e.g., pitch) also has spatial information. Thus, multiple modalities might be mixed when processing auditory spatial information, which cannot be explained by the current form of WMM.

Multiple Resource Theory (MRT)

The multiple resource theory proposes that there are four categorical and dichotomous dimensions that account for variance in time-sharing performance (Wickens, 2002). Each dimension has two discrete levels. If other conditions are equal (i.e., equal resource demand or single task difficulty), two tasks that demand *separate* levels on the dimension (e.g., one visual and one auditory tasks) will interfere with each other *less* than two tasks that both demand one level of a given dimension (e.g., two tasks demanding visual perception). The four dimensions (Figure 2) include processing stages, perceptual modalities, visual channels, and processing codes. Moreover, all of these dichotomies can be associated with distinct physiological mechanisms. Since MRT distinguishes codes from modalities, spatial information can also be processed in auditory modality, which is difficult to be explained that way in Baddeley's WMM. Note that Wickens has also elaborated visual modality more than auditory modality. For example, MRT classifies visual modality into focal vision and ambient (or peripheral) vision, which is also supported by the physiological mechanisms at the eye level – cones (focal) and rods (ambient) and brain level – ventral (focal: what information) and dorsal (ambient: where information) procedures (Yantis, 2014). However, there is little refinement in auditory modality.

Auditory and multimodal display researchers heavily depend on MRT as the primary theoretical backup for use of auditory user interfaces. However, when it comes to nonverbal auditory cues, MRT loses its explanatory power. For example, it is questionable if visual and nonverbal spatial information can be easily time-shared according to MRT. Even more, MRT does not provide a clear explanation if verbal and nonverbal auditory information processing can be easily time-shared as in the visual and verbal information case. These two models (WMM and MRT) are widely accepted and

comprehensive, but of course, not perfect. For example, these models do not account for storage and process of other senses, including smell, touch, and taste.

Musical Working Memory

To close the gap in nonverbal auditory information processing in WMM, researchers have suggested a special working memory model for music or tone processing (Berz, 1995; Pechmann & Mohr, 1992). These models propose that working memory includes a specific part for tonal processing which the traditional WMM does not imply. In the classic WMM, the phonological loop is assumed to process both acoustic and verbal information. Researchers on music working memory attempt to explain music experts' increased performance based on a "tonal loop", instead of visuo-spatial sketchpad or phonological loop. They assert it based on "unattended music effect", which refers to the tendency that unattended speech interferes with the process of linguistic information, whereas unattended music has little effect on it.

In fact, there is some physiological evidence for independent music processing. People have an implicit ability to process both speech and non-speech sounds. The supporting evidence for speech comes from research showing that new born babies are able to make phonetic discriminations (Dehaene-Lambertz & Pena, 2001). Likewise, the evidence for non-speech comes from research showing that infants also detect changes in tonal patterns (Dehaene-Lambertz, 2000). Koelsch and colleagues have also found that adults without formal musical training show that both early (200ms) and late (500ms) ERP responses to unexpected changes in tonal patterns (Koelsch, Gunter, Friederici, & Schröger, 2000). These early ERP component patterns in response to unexpected sound patterns are called "ERAN" (Early Right Anterior Negativity) responses. They proposed that the ERAN reflects a preattentive musical sound expectancy violation based on implicit knowledge of complex musical patterns (Koelsch, Maess, Grossmann, & Friederici, 2003).

Musical working memory certainly advances WMM by specifying the tonal loop. However, it has not been widely accepted or applied in Human Factors. It seems not sufficient to apply to practical settings in several aspects. First, this model explains the music experts' case better than non-experts', which is obviously a rare case. In auditory display literature, many studies have shown that users even without music training significantly benefit from having nonverbal auditory displays (e.g., Brewster, Wright, & Edwards, 1994). Second, not all nonverbal sound is music. The International Community for Auditory Display (ICAD) has a long history in arguing that auditory displays and sonification are very different from music. Of course, there is a type of musical sound like earcons, but other auditory displays, such as auditory icons (natural sounds of the events) and spearcons (compressed speech) are not musical sounds. Obviously, an inference process about all these auditory displays seems to be distinguished from music processing as well as language processing. Therefore, more research is required to validate whether and how performance enhancement by these nonverbal auditory signals can be explained by the musical working memory model. Finally and most importantly, we still need to identify more comprehensive mechanism(s) about representation and encoding processes, rather than simply adding one more special working memory module. Unfortunately, Berz (1995) did not provide detailed explanations about these critical topics of the musical working memory. In conclusion, there is evidence to suggest a new model is necessary to explain these phenomena.

3 PROPOSED MODEL

I propose a preliminary conceptual model describing nonverbal auditory information processing that specifies its representation and encoding process in working memory. Similar to WMM, the nonverbal auditory module has two components: the phonatory store, which holds memory traces in acoustic and nonverbal characteristics (as in a form of sound per se) for a few seconds, and the phonatory rehearsal process, which includes singing or humming. The function of the phonatory rehearsal process is to retrieve and re-articulate the contents held in the phonatory store. In this process, the new information will be matched to phonatory lexicon (e.g., template of timbre) in long-term memory to associate its meanings. However, it is not clear if these two components belong to the phonological loop of the original WMM or an independent new module, "the phonatory loop". In addition to this basic concept, there are multiple variables that influence this process depending on an individual's expertise and long-term working memory.

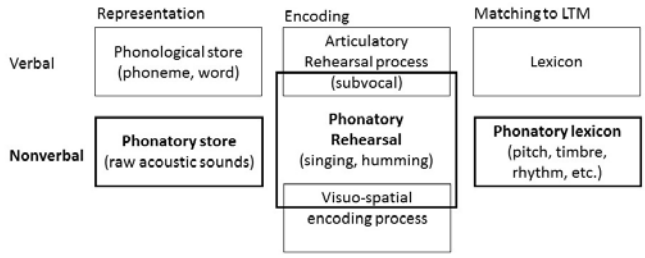

Figure 3. A Conceptual model of verbal and nonverbal auditory information processing. It seems that there are separate representation and encoding processes, but it is not clear whether these distinctive processes happen in the same module (i.e., the phonological loop) or in the different module (i.e., the phonatory loop).

Behavioral Evidence for Representation and Encoding

Research has shown that nonverbal sounds can be rehearsed (Keller, Cowan, & Saults, 1995) and this

internal representation has been referred to as "auditory imagery" (e.g., Brodsky, Henik, Rubinstein, & Zorman, 2003). Here are a couple of supporting studies. In successive studies, Brodsky and colleagues (Brodsky et al., 2003; Brodsky, Kessler, Rubinstein, Ginsborg, & Henik, 2008) investigated the mental representation of music notation using an "embedded melody" task (or pattern matching), in which they embellished the musical phrase using a compositional technique. In this paradigm, participants were asked to silently read the score of an embedded melody. Then, the score was removed from their sight and they were asked to accept or reject if the tune is the same as the original. Results showed that phonatory interference (e.g., involving wordless singing or humming a folk song aloud) impaired recognition of original themes more than did the other conditions (e.g., rhythmic distraction and obstruction by auditory stimuli).

On the other hand, Nees (2009) showed that people can also encode nonverbal auditory information using either verbal or visual imagery, in addition to auditory imagery (i.e., phonatory representation). This suggests that WMM or MRT can still partly explain the encoding process of nonverbal auditory signals. However, in his experiment, the auditory imagery condition showed the shortest mental scanning time. Also, subjective rating scores showed lower perceived workload for the auditory imagery condition than the other two conditions because the external stimulus did not have to be *recoded* into a different internal format. Furthermore, his second experiment showed that the generation of a visuospatial image from a sound took longer than the generation of an auditory imagery. Overall subjective workload was highest under the visuospatial imagery encoding condition, which naturally suggests that humans might not likely visually process auditory signals even though they can do if they are asked.

There might also be some processing differences between experts and novices. Pechmann and Mohr (1992) revealed clear group differences between musicians and non-musicians. In Deutsch's experimental paradigm (i.e., target tone – 6 distractors – test tone), musically trained subjects' retention of the first test tone was only affected by the interposition of other tones. In contrast, the performance of musically untrained subjects was also affected by verbal and visual distractors. This study supports that non-music experts try to use other strategies (phonological loop or visuo-spatial sketchpad).

In addition to using different representations in working memory, experts' performance is also affected on the strategic level. For example, experts are known to have a quick access to long-term working memory. Their strategies include a schemata (Berz, 1995) about patterns or chunking. Jeon and Hahn (2003) compared working memory processing in music reading tasks between music experts and novices. Participants were asked to remember musical notes (either melodious or unmelodious sequence in experiment 1; either harmonious or disharmonious chord sequences in experiment 2) on the screen. Participants had to encode the notes either with or without the articulatory suppression (i.e., they were asked to repeat "da da da da" while they encode the notes.). Results showed consistent better performance in the expert group than in the novice group and better performance in the non-articulatory suppression condition than the articulatory suppression condition with both stimuli types. However, experts were less affected by the articulatory suppression when the stimuli were melodious or harmonious than the other conditions. These findings support the hypothesis that expert musicians can use different strategies in music reading in addition to the differences in working memory. They might chunk melodious lines in experiment 1 and use their schemata of familiar chord progressions in experiment 2.

Neurological Evidence for Representation and Encoding

In addition to preattentive ERAN responses based on musical sound expectancy (Koelsch et al., 2003), our brain has a number of separate mechanisms about nonverbal auditory information processing (e.g., pitch, loudness, time difference, auditory localization, to name a few). In fact, many books about auditory perception and psychoacoustics allocate most chapters to this acoustic information processing. Speech processing is usually just one chapter out of the entire book (e.g., Moore, 2012). As a representative example, to process pitch perception, our basilar membrane has a tonotopic arrangement of frequencies, which is grouped into 24-30 critical bands. This tonotopic mapping is preserved across auditory nerve fibers and auditory cortex. This is specifically designed for frequency processing, which is unique and distinctive from speech processing.

On the other hand, there is research to show that the topography of activations was virtually identical for the rehearsal of syllables and pitches (Koelsch et al., 2009), which might imply that the same cortical area is used for both rehearsal process of the phonological loop and the phonatory loop. However, certainly different cortical areas are involved in language perception from pitch perception. Language perception heavily relies on temporal changes in broadband sounds, but music perception depends more on the ability to discriminate slower, more precise changes in frequency (Zatorre, Belin, & Penhume, 2002). Temporal resolution is better in the left hemisphere (which is responsible for more language processing), whereas the right hemisphere more specialized for spectral resolution (which is essential to music perception). Over the course of three experiments Williamson, Baddeley, and Hitch (2010) showed that

short-term memory for tones has several similarities to verbal memory, including limited capacity and a significant effect of pitch proximity in non-musicians. Despite being vulnerable to phonological similarity when recalling letters, musicians showed no effect of pitch proximity, which suggests again there might be fundamental differences between verbal processing and nonverbal processing for trained musicians.

Evaluations of the Proposed Model and Future Works

Requirements must be met to scientifically validate a new construct in cognitive engineering (e.g., Parasuraman, Sheridan, & Wickens, 2008). First, it should be distinguished from the existing construct to contribute to a novel area which the previous one cannot explain. Simultaneously, it should be linked to the existing construct. Second, there should be underlying neurological mechanisms. Finally, the construct should be modeled computationally to simulate and predict plausible behavioral outcomes. When all these criteria have been satisfied, converging evidence for the scientific validity of the construct will arise. I believe that the proposed model describing nonverbal auditory information processing can account for many phenomena regarding multimodal user interfaces (e.g., when earcons, auditory icons, and spearcons are used with graphical user interfaces), which cannot be fully explained by the existing models. As briefly discussed, there is much neurological evidence suggesting separate processing of acoustic and nonverbal auditory information at a neurophysiological level. However, it remains unanswered whether verbal information and nonverbal information are processed in the same module (phonological loop vs. phonatory loop). In terms of time sharing, we can use an analogy of vision. Focal vision and ambient vision are separate processes. However, within the same modality, they might not be well-time shared as much as multimodality. Likewise, even though verbal and nonverbal processes are separate, there might be observable interference as long as they have overlapping activation areas. There is not much research on mathematical or computational modeling of nonverbal auditory information processing. We have just started to construct a computational model that can predict driver performance when having nonverbal auditory cues compared to the no auditory condition. I hope that the introduction of this model could help researchers discuss underlying mechanisms of nonverbal auditory processing and develop a valid model about it to predict and explain user performance of multi-tasking in applied settings.

REFERENCES

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556-559.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of psychology, 63*, 1-29.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation, 8*, 47-89.

Berz, W. L. (1995). Working memory in music: A theoretical model. *Music Perception*, 353-364.

Brewster, S. A., Wright, P. C., & Edwards, A. D. (1994). *A detailed investigation into the effectiveness of earcons.* Proceedings of the Santa Fe Institute Studies In The Sciences Of Complexity.

Brodsky, W., Henik, A., Rubinstein, B.-S., & Zorman, M. (2003). Auditory imagery from musical notation in expert musicians. *Perception & Psychophysics, 65*(4), 602-612.

Brodsky, W., Kessler, Y., Rubinstein, B.-S., Ginsborg, J., & Henik, A. (2008). The mental representation of music notation: notational audiation. *Journal of Experimental Psychology: Human Perception and Performance, 34*(2), 427.

Dehaene-Lambertz, G. (2000). Cerebral specialization for speech and non-speech stimuli in infants. *Cognitive Neuroscience, Journal of, 12*(3), 449-460.

Dehaene-Lambertz, G., & Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport, 12*(14), 3155-3158.

Jeon, M., & Han, K. H. (2003). *Difference in processing type of working memory in music reading between experts and novices.* Proceedings of the 4th International Conference of Cognitive Science, 222-227, Sidney, Australia, July, 2003.

Keller, T. A., Cowan, N., & Saults, J. S. (1995). Can auditory memory for tone pitch be rehearsed? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(3), 635.

Koelsch, S., Gunter, T. C., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing:“nonmusicians” are musical. *Cognitive Neuroscience, Journal of, 12*(3), 520-541.

Koelsch, S., Maess, B., Grossmann, T., & Friederici, A. D. (2003). Electric brain responses reveal gender differences in music processing. *Neuroreport, 14*(5), 709-713.

Koelsch, S., Schulze, K., Sammler, D., Fritz, T., Müller, K., & Gruber, O. (2009). Functional architecture of verbal and tonal working memory: an FMRI study. *Human brain mapping, 30*(3), 859-873.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Nees, M. A. (2009). Internal representations of auditory frequency: behavioral studies of format and malleability by instructions. PhD Dissertation, Georgia Institute of Technology, GA.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160.

Pechmann, T., & Mohr, G. (1992). Interference in memory for tonal pitch: Implications for a working-memory model. *Memory & Cognition, 20*(3), 314-320.

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological bulletin, 113*(2), 345.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*: 3rd Edition. Prentice Hall.

Wickens, C. D., & McCarley, J. S. (2007). *Applied attention theory*. CRC press.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2006). *Music in working memory? Examining the effect of pitch proximity on the recall performance of nonmusicians.*Proceedings of the 9th International Conference on Music Perception and Cognition.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition, 38*(2), 163-175.

Yantis, S. (2014). *Sensation and Perception*. NY: Worth Publishers.