One singer, many voices: Distinctive within-singer groupings in Tom Waits

Joshua Albrecht
Hugh A. Glauser School of Music, Kent State University, Kent, OH, USA
jalbrec6@kent.edu

Introduction

Timbre plays an essential role in popular music (Fink *et al.,* 2018, p. 2), in which a unique sound can craft a distinctive identity for a band or artist even as other musical features show less variability, such as stock chord progressions and more limited melodies. A singer's vocal quality alone can create a recognizable persona (Tagg, 2012, p. 350) and become associated with the singer's brand. Moreover, the human voice is one of the most sophisticated and musically relevant sources of timbral variety in music, and humans are remarkably adept at deciphering timbral cues in voices. Hughes *et al.* (2004) found that listeners are consistently accurate at inferring speakers' age, height, weight, socio-economic status, personality traits, and emotional and mental states from listening to recordings alone. These results suggest that one may be able to define "sonic fingerprints" for individual singers by examining distinctive acoustic characteristics of their voice, at least in theory. However, pinpointing the sonic markers that distinguish one singer from another by examining only the audio source can be difficult, and the acoustic parameters associated with recordings of different singers are likely confounded by many other recording artifacts.

An alternative approach would control for *singer* and focus on different timbral approaches the same singer uses to influence voice quality. Several singers are known for mastery over a wide range of vocal timbres, such as Billy Joel (Duchan, 2016) and Bob Dylan (Rings, 2013). One of the more extreme cases is that of Tom Waits. One of the most immediately recognizable aspects of his music is his distinctly rough vocal timbre(s). Solis (2007) argues that Waits actually uses many different voices, each distinct and recognizable. While rock music since the 1960's typically claims some form of implicit autobiography, Waits' songs by contrast are (often overtly) inhabited by fictional personas who speak in distinct voices, both figuratively (lyrical meaning) and literally (unique vocal timbres). However, many of his songs share similar vocal timbres, suggesting links or shared meaning between them. The music of Tom Waits provides an interesting case study to compare groups of recordings that are perceived to have similar vocal timbres against other groups that are perceived as having different timbres, but yet controlling for possible confounds associated with different singers.

Method

The ideal approach for a study like this would be to locate masters of recordings and isolate the voice. Vocal tracks could then be subjected to automatic feature extraction and/or a perceptual study. Unfortunately, all of my attempts to contact Waits's studios have been unsuccessful, and so all tracks examined consisted of Waits's voice along with instrumental accompaniment. Without being able to isolate the voice, automatic feature extraction would be problematic and similarities between recordings would be likely to be strongly influenced by instrumentation rather than vocal quality alone. However, human listeners are exceptionally skilled at auditory scene analysis and can sift timbral properties of voices out of complex acoustical environments (Bregman, 1990). Consequently, this study used a perceptual method with human participants who listened to excerpts and sorted them by hand. Due to the important role that semantic description plays in capturing perceptually relevant timbral properties (Saitis & Weinzierl, 2019), participants also provided descriptive labels for their timbral categories

Sample:
Tom Waits's entire output, consisting of 255 vocal tracks over 19 studio albums, was too large for the present study. However, his career is often divided into two phases. The second phase, beginning after Waits married his wife Kathleen Brennan who encouraged him to experiment with a more adventurous

range of vocal timbres and instruments, is more relevant to the current study. Beginning with the album *Swordfishtrombones* (1983), Waits produced 146 tracks with voice over ten albums. For this study, the first five seconds that included at least 4 seconds of vocal sound from each of these 146 tracks was sampled.

Participants:

A total of 134 undergraduate music majors participated, 73 from the University of Mary Hardin-Baylor and 61 from Kent State University. 84 of the participants were female, and 50 were male, with a mean age of 20.7 (sd = 5.5). 101 participants reported Rock music as the musical genre they primarily listened to, with 21 reporting classical music, 11 jazz, and 1 not reporting. Discerning timbral differences is a challenging skill, and so musical sophistication was an important consideration for participant selection. The majority of participants were music major undergraduates, and participants averaged a Goldsmith's Musical Sophistication Index for general musical sophistication of 97.3 out of 126 total (sd = 11.7). Average scores of subsets of the Gold-MSI measure included 47.8 out of 63 for active engagement (sd = 6.9), 51.0 out of 63 for perceptual abilities (sd = 6.1), 36.7 out of 49 for musical training (sd = 6.8), 34.9 out of 42 for emotions (sd = 4.7), and 35.5 out of 49 for singing abilities (sd = 6.4).

Procedure:

Participants were provided 40 randomly-selected five second excerpts from the full set, represented as boxes on a computer screen (see Figure 1). The interface was run on a private Amazon Web Services website through the Google Chrome browser. UMHB participants were seated in groups of 8 in a computer lab using headphones. Due to the COVID-19 pandemic, KSU participants used their personal computers, but were instructed to complete the experiment in one sitting in a distraction-free environment, using headphones. After listening to their excerpts in numerical order, participants could drag the boxes into as many as eight groups or categories and re-listen as many times as needed by clicking on the box's play button. Participants were told to sort the excerpts according to vocal timbre and to ignore instrumentation, texture, meaning of the text, and genre. After the initial sort, participants were presented with each of their assembled groups in the second phase, listening to all excerpts in the group in random order. They were then asked to provide the best description of the vocal timbre of the excerpts in the group. In the third phase, participants progressively merged groups until there were only two left.

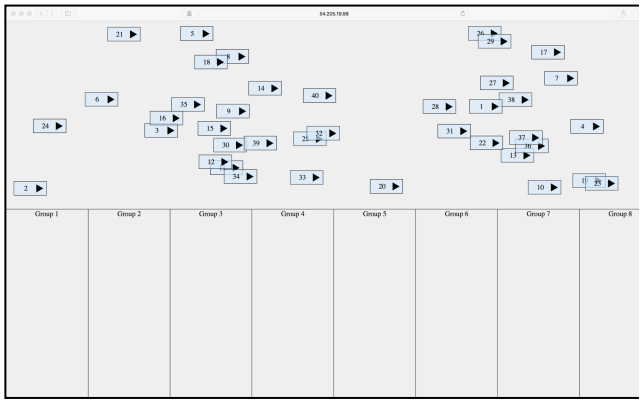

Figure 1: The interface. 40 randomly-selected excerpts scattered across the top of the screen and participants freely dragged them into as many as eight categories.

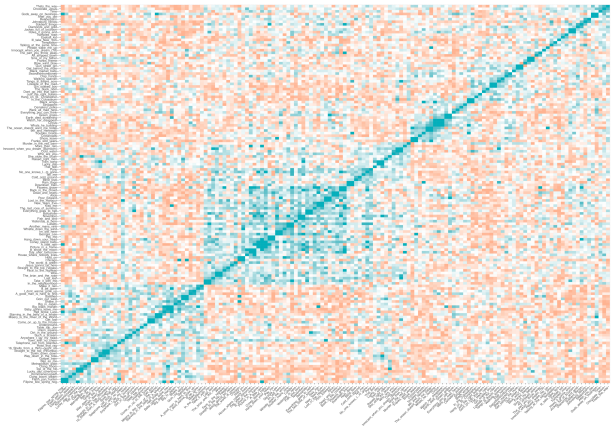

Figure 2: Dissimilarity matrix of the participant grouping responses for the entire dataset. Max dissimilarity is in red and minimum in blue.

Results

Participant data were used to estimate the timbral similarity between each pair of excerpts. Similarity measures were obtained by calculating the number of times two excerpts were grouped together at any stage of the grouping procedure for any participant by the number of times they *could have been* grouped

together. For example, if two excerpts were always grouped together every time they were co-present, they would have a similarity of 1, and if they were never grouped together in any condition, they would have a similarity of 0. Every possible two-excerpt pair were presented together in the same experiment at least four times. The grouping data were subjected to cluster analysis, in which the proportion of excerpt grouping was treated as the distance measure. Cluster analysis require *dis*similarity data, so the similarity scores were subtracted from 1. In other words, if two excerpts were never grouped together, their dissimilarity was 1. The dissimilarity matrix for preliminary results on data from the 72 UMHB participants for all 146 excerpts appears in Figure 2. Maximum dissimilarity is shown in orange and minimum dissimilarity is shown in blue.

It is not straightforward to determine the optimal number of clusters in a cluster analysis. A number of metrics have been proposed to provide an empirical means of determining the optimal number of clusters, though the intuition of the research remains an important consideration. The gap statistic is a test of how many clusters are most appropriate in a given dataset by comparing the total within-cluster variation for different numbers of clusters against the expected values under null reference distribution of the data (Tibshirani *et. al* 2001). The gap statistic for the preliminary dataset suggests an optimal number of seven clusters (see Figure 3). Hierarchical clustering dendrograms are also useful to visualize how many clusters appear appropriate for the data. A dendrogram showing Ward's method presented in circular form for reasons of space appears in Figure 4 with the 7-cluster solution shown. This solution appears robust, with only a 3-cluster solution appearing to have more distance between cluster heights. After verifying the appropriateness of 7 clusters, a separate k=7 k-means cluster analysis was conducted. A 3D multidimensional scaling of the results are provided in Figure 5. Clustering is apparent, but weak.

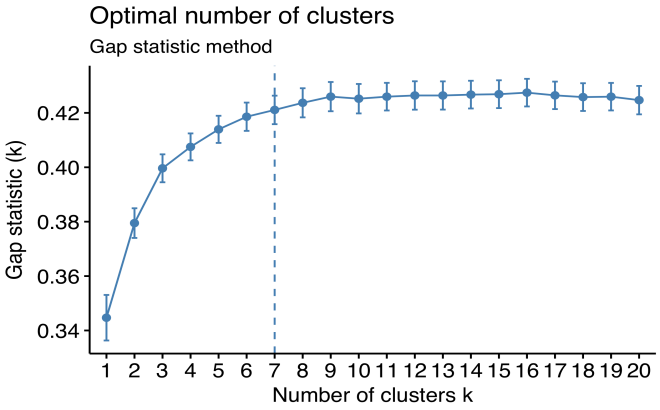

Figure 3: Dissimilarity matrix for the participant grouping responses. Maximum dissimilarity is in red.

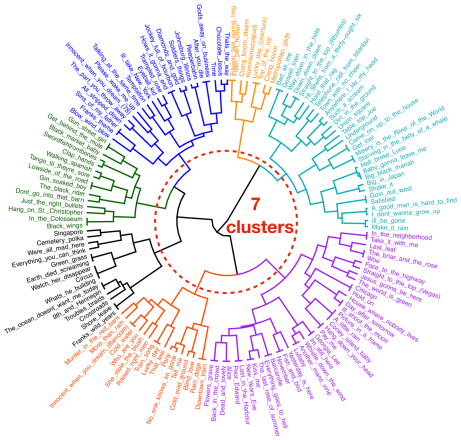

Figure 4: Ward's hierarchical cluster results for seven clusters

Once initial groups were established, each participant provided a descriptive label of the timbral quality of each excerpt as a free response. These responses were cleaned by removing all words that were not descriptors (generally, adjectives were kept and other words removed). When appropriate, nouns were turned into adjectives, and adjective modifiers like "really" or "very" were removed. When there were two adjectives in different forms, the simpler was retained, so "stranger" was changed to "strange." After cleaning the timbral descriptors, there remained 6,124 total terms and 407 unique terms describing Waits's vocal timbres of the preliminary dataset. The seven most common terms for each of the seven k=7 k-means clusters are displayed in Table 1. Of note is that "raspy" is the most common descriptor for all seven clusters, appearing a remarkable 9.5% of the total dataset. But, the rank-ordering of the next most common terms for each cluster reveals interesting differences between patterns of perceptions of vocal timbre for each cluster. For example, after raspy, C1 is described as deep, smooth, and breathy, whereas C6 is rough, growly, screaming, and harsh, and C4 is speech-like, deep, and low. The rank-ordered differences are consistent with different timbral characteristics for each cluster.

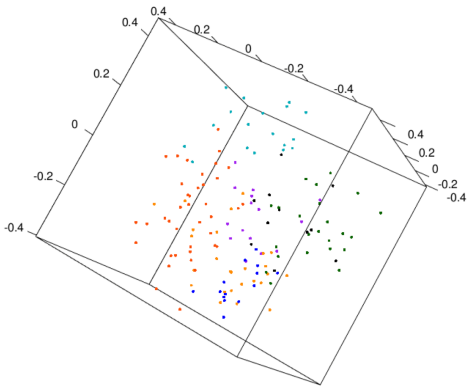

Figure 6: 3D multi-dimensional scaling for k=7 k-means clustered data.

C1 (33)	C2 (22)	C3 (15)	C4 (21)	C5 (16)	C6 (16)	C7 (23)
raspy (150)	raspy (83)	raspy (48)	raspy (93)	raspy (58)	raspy (52)	raspy (97)
deep (76)	nasal (45)	smooth (35)	speech-like (42)	rough (18)	rough (31)	growly (38)
smooth (74)	rough (29)	breathy (31)	deep (33)	deep (15)	growly (25)	rough (38)
breathy (63)	harsh (19)	soft (21)	low (29)	smooth (15)	scream-ing (21)	deep (27)
soft (45)	speech-like(19)	relaxed (16)	soft (27)	breathy (14)	harsh (15)	yelling (22)
low (38)	jazzy (18)	light (15)	breathy (24)	growly (14)	nasal (14)	gravelly (20)
scratchy (30)	light/ growly(13)	airy (11)	smooth (24)	heavy (13)	scratchy (13)	husky (17)

Table 1: The most common timbral description terms by cluster.

Discussion

Participant grouping data for the corpus of Tom Waits's studio songs is consistent with the existence of a small number of discernible groups based on the vocal timbre used, suggesting that this dataset could be a useful tool for examining inter-singer timbral differences. The qualitative descriptions provided by participants reveals striking differences between groups from the same singer. While the preliminary data suggests that seven groups is an appropriate number of song subsets in the Waits corpus, more refined subsets could be possible when the additional 61 participant dataset is added to the analysis. A close examination of Figure 2 reveals the presence of dozens of small clusters of 3-5 songs that are highly similar.

References

Bregman, A. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: The MIT Press.

Duchan, J.S. (2016). Depicting the working class in the music of Billy Joel. In K. Williams & J. A. Williams (eds.), *The Cambridge Companion to the singer-songwriter*. Cambridge: Cambridge University Press.

Fink, R., Latour, M., & Wallmark, Z. (2018). *Timbre in popular music: The relentless pursuit of tone*. Oxford: Oxford University Press.

Hughes, S.M., Dispenz, F., Gallup, G.G. (2004). "Ratings of voice attractiveness predict sexual behavior." *Evolution and Human Behavior*, *25*, 295-304.

Montandon, M. (ed.) (2005). *Innocent when you dream: The Tom Waits Reader.* New York: Carroll & Graf.

Rings, S. (2013). A foreign sound to your ear: Bob Dylan performs "It's alright, ma (I'm only bleeding)," 1964-2009. *Music Theory Online*, *19* (4).

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In K. Seidenberg, C. Saitis, S. McAdams, A. Popper, & R. Fay (eds.), *Timbre: Acoustics, Perception, and Cognition*. (pp. 119–149). Springer Handbook of Auditory Research. Springer, Cham.

Solis, G. (2007). "Workin' hard, hardy workin'/ Hey Man, you know me": Tom Waits, sound, and the theatrics of masculinity. *Journal of Popular Music Studies, 19* (1), 26-58.

Tagg, P. (2012). *Music's meanings: A modern musicology for non-musos.* New York: The Mass Media Music Scholars' Press.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B, 63* (2), 411-423.