Catching the Lyrics: Intelligibility in Twelve Song Genres

Nathaniel Condit-Schultz & David Huron
Ohio State University

Although purely instrumental music is commonplace, much of the world's most popular music is sung with lyrics. However, it is evident that listeners don't always attend to lyrics and that those who do aren't always successful in deciphering them. An empirical study is reported whose goal is to measure the intelligibility of lyrics in commercial recordings of music from a variety of genres. Thirty participants were exposed to 120 brief musical excerpts from twelve song genres: Avante-garde, Blues, Classical, Country, Folk, Jazz, Musical Theater, Pop/Rock, Rhythm and Blues, Rap, Reggae, and Religious. Participants were instructed to transcribe the lyrics after hearing each excerpt once. The transcribed lyrics were then compared to the actual lyrics and intelligibility scores calculated. The different genres were found to exhibit significantly different levels of lyric intelligibility, from as low as 48% for Classical music, to as high as 96% for Jazz, with an overall average of 72%. Intelligibility scores were positively correlated with listener judgments of the general importance of lyrics. In a second experiment, participants were allowed to hear excerpts five times. Improvements to intelligibility were modest but significant after the second and third hearings, but not on further hearings.

Received: December 4, 2013, accepted September 25, 2014.

Key words: intelligibility, lyrics, singing, music, genre

Although purely instrumental music is commonplace, much of the world's music includes the use of the human voice. In addition, most vocal music makes use of lyrics in preference to nonlinguistic vocables. The ubiquity of song in musical listening suggests that lyrics are an important part of the musical experience for many listeners. Lyrics and their interrelationship with music have long been the subject of extensive scholarly and casual discussion. However, the extent to which listeners actually attend to the meaning of lyrics is not known, and anecdotal observation suggests that the importance accorded to lyrics may vary between styles and between listeners. An oft neglected, yet crucial, issue regarding lyrics is their intelligibility; even when listeners do attend to lyrics they are often unable to understand them. Anecdotes of misheard lyrics are commonplace, with many (often comical) examples cataloged on several popular websites (e.g., www.kissthisguy.com, www.amIright.com). If listeners cannot decipher lyrics correctly, discussions concerning the role of lyrics in the experience of music become moot. This paper reports the results of two experiments designed to establish baseline measures of how well (or how poorly) listeners are able to decipher sung lyrics in recorded music. We also endeavor to gather preliminary data on interpersonal and interstylistic variation in the perception of lyrics.

Literature

A number of previous studies have examined the intelligibility of sung words. Most work has focused on the discriminability of sung vowels. For example, Smith and Scott (1980) showed that the intelligibility of vowels is significantly reduced when the sung pitch exceeds F5. They further noted that intelligibility depends on singing style. When sung with a raised larynx (as might be done in popular music styles) the intelligibility of high pitched tones was significantly improved compared with the lowered larynx commonly found among classically trained singers. Benolken and Swanson (1990) replicated the work of Smith and Scott, showing that sung vowels become increasingly difficult to discriminate as the fundamental frequency is increased. Similarly, Hollien, Mendes-Schwartz, and Nielsen (2000) found that the intelligibility of sung vowels declines considerably when the fundamental frequency reaches or exceeds the typical first formant, as is often the case for soprano singers. Apart from the difficulties involved in discriminating vowels, other aspects of phonology might be expected to contribute to problems in intelligibility. Burleson (1992) speculated that rhythmic aspects of prosody, such as word stress, might also be disrupted by musical settings. However, Burleson did not offer an empirical demonstration of such disruptions.

Adapting intelligibility measures used in architectural acoustics, Collister and Huron (2008) found that roughly a quarter of all words sung by an unaccompanied soloist were misheard by listeners. Target words were sung as

Music Perception, volume 32, issue 5, pp. 470–483, issn 0730-7829, electronic issn 1533-8312. © 2015 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's rights and permissions website, http://www.ucpressjournals.com/reprintinfo.asp. DOI: 10.1525/mp.2015.32.5.470

the last word/note in a "carrier phrase" ("I am singing the word ________") which was set to a variety of simple musical tunes drawn from the classical tradition. Both Classical and Theatrical vocal styles were tested. Sung words showed a seven-fold decrease in intelligibility compared with their spoken counterparts. An analysis of phonetic errors indicated that common perceptual mistakes included the centralization of vowels as well as confusion among voiced stops and nasals. In a follow-up study, Johnson, Huron, and Collister (2012) identified a number of further sources of confusion, including the presence of melismas (single syllables spanning many pitches), the tendency to employ archaic vocabulary in vocal texts, and the mismatch of stress between words and musical rhythm—all features known to vary with musical style.

AIMS

Previous studies have used solo voices and isolated phonemes or words as stimuli, resulting in controlled conditions with low ecological validity. One can easily imagine how real musical context might either enhance or degrade intelligibility. Instrumental accompaniment, vocal harmonization, reverberation, singing style, and other factors might be expected to interfere with lyric deciphering, while predictable rhyme schemes and semantic context might improve intelligibility. In the current study we examine lyric intelligibility in conditions more closely approximating normal listening situations, specifically the intelligibility of commercial music recordings. Since musical features, contexts, and artistic goals vary with style, one might expect intelligibility to vary with style as well. For example, one might expect that Broadway songs would be easy to understand, since the dramatic story in this genre is typically central to the artistic goals. Accordingly, we resolved to compare intelligibility measures for a number of common musical genres. In addition to this interstylistic variation in the role of lyrics and their intelligibility, we also expect interpersonal variation. Individuals' personal listening habits and preferences certainly vary and we conjecture that this will correlate with their ability and desire to decipher lyrics. We thus gathered personal data from participants to test this conjecture.

Experiment 1
Method

STIMULI

Since the majority of music listening involves commercial sound recordings, it is appropriate to consider commercial recordings to be ecologically valid stimuli. For the purposes of this study we limit ourselves to English-language songs, performed by native speakers of English and heard by native speakers. It is likely that differences of dialect may influence intelligibility. However, certain genres are strongly associated with specific dialectics (for instance Reggae music and Jamaican dialects). Since these accents form a part of the normal listening canon it is appropriate to include these dialects when estimating the intelligibility of vocal music. We aimed to assemble a sample representing the breadth and variety of music common among contemporary English-speaking listeners.

Our first consideration was which musical styles, or genres, to compare. The idea of genre is a complicated subjective phenomenon and we anticipated that no genre categorization would be perfect. However, our aim was not to assemble a sample representative of general music listening but to capture a wide spectrum of vocal styles. Thus, we sought only broad, commonplace, genre distinctions. To minimize experimenter bias we utilized an independent taxonomy to select and distinguish genres: the All Music Guide (www.allmusic.com). The All Music Guide is an extensive online database of sound recordings—including detailed publisher information, reviews, and biographical information—which distinguishes some twenty genres. Eight of these genres were considered inappropriate for our purposes—including Comedy, Children's, Electronic, and Latin music. Twelve were judged to be pertinent to our experiment: Avante-garde, Blues, Classical, Country, Folk, Jazz, Pop/Rock, Rhythm and Blues, Rap, Reggae, Religious, and Stage & Screen (hereafter referred to as "Theater"). Despite the differences in the size, variety, and popularity of these genres we employed a stratified sampling method, sampling equally from each.

Having selected genres, the next step was to sample specific songs. In order to properly test intelligibility we sought excerpts that would be unfamiliar to our participants. We thus avoided especially popular artists and songs. Another independent source, the website Rate Your Music (rateyourmusic.com), was used to operationalize familiarity and make random album selections. Rate Your Music (RYM) is a database of popular music albums organized wiki style by users who contribute reviews and ratings. On RYM, famous albums have thousands of user ratings whereas obscure albums have fewer than fifty ratings. By searching the RYM database by genre and sorting the results by the number of user ratings it was possible to avoid albums with large numbers of ratings, with the expectation that albums with fewer ratings are relatively unknown. RYM proved a useful source for selecting Pop/Rock (though biased towards

Rock), Blues, and Country albums but was less useful for other genres—in the end contributing to the selection of 130 excerpts. For other genres, particularly Avante-garde and Classical, we were forced to resort to less systematic sampling, simply searching AMG or Google to find artists. For the purposes of this study, the main resources for gathering audio recordings were www.rdio.com and grooveshark.com—two licensed online streaming sites. Other excerpts were chosen and acquired from a convenience sample of CDs that could be accessed through the Ohio State University and Columbus Metropolitan Library systems. Library CDs are certainly not a random sample as they are biased towards relatively popular artists and genres (especially Rock and Country). A track number from each CD/album was randomly selected in order to pick a song. In order to achieve a reasonable statistical power we sought to assemble 20 excerpts from each genre. We were ultimately able to meet this goal for all genres except for Avante-garde, for which only 18 suitable excerpts were found. For two genres, an excess was found, 24 for Country and 26 for Pop/Rock, resulting in a total of 248 excerpts.

In order to avoid biasing excerpts to any particular formal part of a piece, a random point in each recording was selected and the nearest sung phrase (either before or after) was selected. Thus, the selected excerpts represent a mixture of formal sections, including verses, choruses, and bridges. Since metric placement may affect intelligibility, excerpts were edited to be preceded by at least two beats of metrical context. The intention was to help the participant orient metrically before the singing begins. By splicing instrumental breaks from elsewhere in the song, it was possible in nearly all excerpts to edit two to four beats of metrical (and tonal) context before the singing began. However, in a few cases there were no points in the song where the vocal parts rested long enough to create an instrumental introduction. Eight excerpts exhibited this problem to some extent, and thus start rather abruptly with little or no context. Nine additional excerpts were solo voice and thus also had no introductory material (two Avante-garde, two Classical, three Folk, and two Religious). In most cases, excerpts were edited to include a short (≈ 1 second) fade-out after the target phrase.

Excerpt length. The length of excerpts is an important consideration, as overly lengthy excerpts might tax short-term memory, confounding our results. Research suggests that humans have the ability to hold approximately seven independent objects (or "chunks") in short-term memory concurrently (Miller, 1956). Thus, a length limit of approximately seven might be appropriate for a random sequence of words. However, research has also shown that semantic and grammatical context allows listeners to recall considerably longer utterances—in the range of 14–16 words—without undue difficulty (Chen & Cowan, 2005; Gilchrist, Cowan, & Naveh-Benjamin, 2008). In a situation where the listener may not be successful at deciphering the words (and thus may have incomplete grammatical or semantic information) it is not obvious what length limit would be appropriate. Another important consideration is that excerpts represent typical musical phrases from a particular genre. Thus, if Rap is characterized by rapid phrases with many words while Classical tends towards drawn out phrases with few words, it is appropriate that our excerpts represent this stylistic variation. Ultimately, excerpts were simply edited to represent a single stylistically coherent linguistic/musical segment, roughly equivalent to a single sentence or clause. The resulting stimuli average 7.5 seconds in duration ($SD = 3.4$) and contain an average of 8.8 syllables ($SD = 2.8$), forming an average of 7.3 words ($SD = 2.3$). Thus, the excerpts average 1.2 syllables per second ($SD = 0.5$), though this is an underestimate of actual pronunciation rate because excerpt lengths included instrumental portions. The longest excerpts entail 14 words, 18 syllables, and 26 seconds respectively. Given the memory constraints suggested by the literature these lengths are not unreasonable. Further discussion of possible memory confounds is presented in the discussion. A total of 248 excerpts were prepared, comprising 1,802 words. Though practical considerations forced us to compromise some of the a priori ideals we set out to achieve, we believe that the gathered excerpts still represent an appropriate sample for this broadly aimed study.

Excerpt texts. In order to test intelligibility, it is necessary to establish the "correct" transcription for each excerpt, representing the words the singer intended. Thus, we initially sought "official" printed versions of lyrics, presumably written or at least approved of by the performer. We searched for lyrics published in CD packaging along with recordings or posted on websites maintained by the artist or their publishing company. Unfortunately, finding "official" lyrics proved to be challenging; locating official lyrics online proved possible for just two of the final excerpts. Lyrics printed in CD packaging provided "official" lyrics for 77 excerpts. For Theater and Classical excerpts, printed lyrics were readily available in librettos and scores, providing an additional 35 "official" lyrics. For the remaining excerpts we were forced to rely on unofficial lyrics. One-hundred and nine were copied from unofficial online lyrics sites. Since these sites have no official status some degree of mistrust is warranted. Nearly all of these sites allow users to correct or edit the lyrics so there is a wiki-like community effect that may ensure a degree of accuracy not otherwise expected. We compared several different websites for the same lyrics and found that there is a high degree of concordance. However, lyrics at one site may have simply been copied from another site, so independent web sites do not guarantee independent transcriptions. For the final 27 excerpts the lyrics were simply transcribed by the experimenter, as (mostly in the case of Avante-garde and Religious works) it was not possible to locate any independent transcriptions of the lyrics. During the editing process it became clear that even some of the "official" lyrics differed slightly from what was sung on the recording. In order to better validate the acquired transcriptions, six auditors were recruited to provide independent assessments of transcriptions' accuracy. Auditors received the following instructions:

> Listen to each excerpt at least twice, until you feel confident about your interpretation of the lyrics. Having formed an impression of the lyrics, then look at the printed transcription and identify any deviations from your own interpretation. Provide a count of the number of deviations from your interpretation to the transcribed interpretation. If there's a deviation, listen to the passage at least twice again, and resolve in your mind whether you think your original interpretation or the transcribed interpretation is more likely to represent the true lyrics.

Twenty-seven excerpts had words in the collected lyric that the experimenter judged were either clearly missing, or inaccurately included. For an example of a missing word, the official CD booklet for the Blues excerpt from the song "Talk to Ya" had the lyrics as "You're so far away" but on the recording the singer clearly adds the word "baby" after "away." Thus, the lyric was presented to the auditor as "you're so far away (baby)." For an example of a word added to the lyrics, the official CD booklet for the country excerpt from the song "Born Lonesome" reads "the wind blew cold on that mountain," yet the singer clearly does not sing the word "the." Thus, the lyric was presented to the auditor as "(The) wind blew cold on that mountain." Auditors were given additional instructions to judge all 27 parenthetical words:

> Some of the transcribed lyrics contain words in (). For any of these excerpts, please type in column D what you think the parenthetical word is (in some cases you might think the parenthetical word simply is not there).

Each of the six auditors listened to 83 of the 248 stimuli; consequently there were two independent assessments of each stimulus. A priori, we resolved to attend to only those discrepancies to which both the auditors agreed. The six auditors identified 31 discrepancies between their interpretation and the transcribed lyrics. Twenty-one of the thirty-one discrepancies were extremely minor, involving a single unimportant word or affix, such as "she" vs. "she's" or "then" vs. "and." Of the 31 discrepancies, 28 were discrepancies identified by only one of the two auditors. This left three discrepancies in which both auditors disagreed with the transcribed lyrics. One of the three discrepancies identified by two auditors originated in an "official" source—as did ten out of the 28 discrepancies identified by one auditor—and thus was not discarded. Another double-auditor discrepancy involved only a single unimportant word; this word was excluded from analysis but the remainder of the excerpt was included. Finally, the last doubly discrepant excerpt was excluded from the analysis completely. Regarding the parenthetical lyrics, all auditors agreed with the experimenter's judgments. Ultimately, we must recognize that there is no perfect way of establishing what the "actual" lyrics are for a song. This conundrum is not unique to our study: Establishing ground truth for comparative studies is often difficult in various areas of research. However, the 31 discrepancies discussed here represent a small proportion of the 1,802 words found in the stimulus lyrics. Although this assessment procedure is not infallible, it is safe to assume that the remaining lyrics represent reasonable interpretations of the true lyrics.

PARTICIPANTS

Thirty-two participants were recruited for the experiment. Participants were drawn from two convenience populations, one consisting of sophomore music students participating in an experimental subject pool, and the second consisting of speech and hearing students participating to receive extra credit in a psychoacoustics course. These participants may not reflect well the general population of music listeners. Because of their greater musical experience it is possible that music students may be more adept in "catching" the lyrics than nonmusicians. We were particularly concerned that singers might be especially skilled at deciphering lyrics. In order to guard against this possibility data on each participant's singing experience was gathered (as described below). Conversely, music students might be more attentive to the melodic, harmonic and other aspects of music, and so may be less attentive to lyrics than the general population. Speech and hearing students may also be unusually adept at interpreting speech sounds, musical or otherwise. A hearing screening test was administered to each participant (details below) leading us to discard one participant's data. Another participant's data was compromised due to computer error. Data from thirty participants was ultimately analyzed.

PROCEDURE

Participants were tested individually in a sound attenuated room, listening to stimuli via loudspeakers—two participants listened via headphones in an adjacent office (due to scheduling conflicts). The volume was adjusted to a comfortable listening level for each participant. Participants sat in front of a computer. The following instructions appeared on the screen, and were read aloud and explained by the experimenter:

> The purpose of our study is to identify how well listeners are able to decipher the lyrics in sung music. In this experiment you'll hear a number of short musical excerpts from a wide variety of songs. For each excerpt we'd like you to write down the lyrics as best you can.
>
> After each music example a text box will appear on the screen where you can type the lyrics you heard. You can take as much time as you like to write down the lyrics, but you'll only get one chance to hear each excerpt.
>
> You will probably have never heard any of the examples before. If you do recognize the song or the singer, please still transcribe the lyrics you heard but please also write in parenthesis: "I recognize this song" or "I recognize this singer."
>
> After typing your response, press the ENTER key to continue with the next excerpt. When you're ready to begin press any key.

Each participant listened to and transcribed lyrics for 120 excerpts (ten randomly selected from each genre, in random order). Listeners were not able to type while listening. This was done to prevent possible interference due to the sound of typing, and to avoid inattentiveness during the latter part of the stimulus. Clearly, the experimental task encourages participants to attend carefully to the lyrics. Presumably, not all listeners are so attentive during typical listening situations. Accordingly, as part of our study, we collected basic information regarding normal listening habits related to lyrics. Specifically, we collected two pieces of information (as described below) relating to this issue: one item was collected using a spontaneous listening task, and a second item was collected using a questionnaire.

The experiment described above actually occurred following the spontaneous listening task. After the participant was seated for the experiment, and prior to the instructions, the experimenter said "Let me just check that the sound is working." A randomly selected single musical stimulus was played (which was not an excerpt the participant would hear during the main experiment). Immediately following this, the experimenter asked "Could you repeat the words you just heard?—This is actually part of the experiment." In this case, the experimenter wrote down what the listener reported (rather than the participant typing their response). As might be expected in the disorienting moments prior to an experiment, this unanticipated procedure was not always successful. Consequently, we were unable to collect data for five participants, leaving only data for twenty-five participants. After this single spontaneous trial, the experiment continued as described above.

Following the main experiment, each participant was asked to complete two brief surveys. The first survey appeared on the computer screen immediately after the main experiment ended, and consisted of the following fifteen questions, each employing a seven-or six-point scale with the extremes marked as indicated below.

- "On a scale of one to seven tell us: When you normally listen to music, how attentive are you to the lyrics?"
  "Very attentive"—"Not at all attentive"
- "On a scale of one to seven tell us: How important are the lyrics to you?"
  "Not at all important"—"Very important"
- "Which best describes you?:"
  "Nonsinger"—"Music-Loving Nonsinger"—
  "Amateur Singer"—"Serious Amateur Singer"—
  "Semiprofessional Singer"—"Professional Singer"
- For each of the 12 genres: "How familiar are you with the genre _____?"
  "Very familiar"—"Not at all familiar"

The second survey consisted of an eight-item hearing-screening instrument developed by Corren and Hakstian (1992), which was given to the participant on paper after they were debriefed. The questionnaire estimates possible hearing loss through a series of questions related to everyday listening such as "Can you follow the conversation when you are at a large dinner table? never-seldom-often-frequently-always." The Corren and Hakstian questionnaire has been cross-validated by correlating responses to actual audiometric examinations. The survey authors' recommended cut-off score for "normal" versus "impaired" hearing of 27 was used. Participants who scored lower than this value were excluded from the experiment. Using this exclusion criterion one participant's data was discarded.

Analyses

CODING

Forty-six responses were discarded because the participant indicated that they recognized the artist or song. Due to computer error, an additional 20 responses from one participant were discarded. Thus, a final total of 3,504 responses were coded. The texts transcribed by participants were compared to the validated lyrics for each excerpt. Transcribed words that matched the validated lyrics were counted as correct. Words in the validated lyrics that were not identified or were wrongly identified were counted as incorrect. Homonyms were not considered an error and incorrect spelling was ignored. In addition to the 31 minor discrepancies identified during the validation process (discussed earlier) a few additional discrepancies arose during the coding process, requiring several post hoc coding decisions. In all cases, before analysis it was decided which words in each excerpt were essential to a correct score and which were not. Colloquial spellings such as "runnin'" in place of "running" were considered homonyms and missing word repetitions (15 excerpts), backing vocals (5 excerpts), or expressive vocables such as "oh" or "ah" (4 excerpts) were not considered errors. For instance, points were not deducted for leaving out the repetition of the word "please" in the lyric "Won't you please please please accept my love." In some cases, words incorrectly ordered or extra words were encountered. These cases were scored according to the best judgment of the experimenter. For example, one participant responded "short white skirt, short short white slip shirt" where the validated lyric was "a short plaid skirt, a white short short sleeve shirt." The second occurrence of the word "white" was counted as correct even though the participant incorrectly placed it after the word "short."

RESULTS

Genre comparisons. The thirty participants listened to a combined total of 25,408 words, correctly identifying 18,223 for an overall proportion of correct responses of 71.7%. As expected, there was variation across genres, with observed proportions of correct responses ranging from a low of 47.7% for Classical excerpts to a high of 90.8% for Jazz excerpts. To get more accurate estimates of success rates across genres, a mixed-effects logistic regression model, using maximum likelihood estimation, was constructed using the R environment for statistical computing (version 3.0.3) and the lme4 package (version 1.0-4). Genre was the principal fixed effect of interest. Since simple estimates of success rates for each genre were desired, the model was specified with no global intercept. This approach is an alternative to the standard treatment or sum contrasts codings used to represent categorical variables in regression analyses. The resulting model treats zero as the intercept so that regression coefficients and *p*-values for each level of the factor (each genre) represent the difference from zero, which in the logit space used in logistic regression actually represents a 50% chance of success. Since we anticipated that excerpt length might be a confounding effect, we included the number of words and log-transformed length in seconds of each excerpt as fixed effect covariates. These two measures of excerpt length are not correlated ($r = .05$). Both length covariates were mean centered so that genre success estimates in the model would reflect estimates for average length excerpts. Random effect intercepts for each excerpt were included in the model to account for variation in the intelligibility of excerpts within genres. Random effect intercepts for each participant were also included, accounting for the varying abilities of participants. Finally, a random slope for genre within participants was included, allowing for the possibility that different participants would be affected differently by genre differences.

Log-likelihood comparison tests found genre and log-second excerpt length to be significant fixed effects, while excerpt length in words was not significant (Table 1). χ^2 values in Table 1 are from log-likelihood comparison tests. Slope, standard errors, and *z*-values are from model output and are raw logit values (logit values are also translated to percentages for genre estimates). *Z*-values and associated *p*-values compare each genre estimate to the logit value of zero, which corresponds to a 50% chance of success. The model's estimates for each genre, with 95% and 80% confidence limits, are presented in Figure 1. These estimates represent the expected probability of correctly identifying a word in an excerpt from a given genre. Model estimates vary widely from below 50% (Classical) to above 90% (Theater, Country, and Jazz). Since detailed pairwise comparisons between all genres is not an important goal in the current study, relatively anticonservative 80% confidence intervals are included in Figure 1, in order to encourage flexible interpretation. For instance,

TABLE 1. Fixed Effects in Model 1

Factor	χ^2	*df*	*p*	*Slope (logit / %)*	*SE*	*z*	*p*
Length in log-seconds	4.2	1	.04	−0.66	0.32	−2.06	.04
Length in words	2.7	1	.10	−0.08	0.05	−1.60	.11
Genre	114	12	< .01				
Avante-garde				0.74 / 68%	0.44	1.70	.09
Blues				1.48 / 81%	0.40	3.65	< .01
Classical				−0.09 / 48%	0.40	−0.22	.83
Country				2.61 / 93%	0.38	6.89	< .01
Folk				1.40 / 80%	0.39	3.63	< .01
Jazz				3.07 / 96%	0.43	7.21	< .01
Pop/Rock				0.71 / 67%	0.36	1.96	.05
Rap				0.77 / 68%	0.41	1.88	.06
R&B				1.63 / 84%	0.40	4.08	< .01
Reggae				1.35 / 79%	0.40	3.37	< .01
Religious				0.97 / 72%	0.41	2.34	.02
Theater				2.42 / 92%	0.41	5.86	< .01

TABLE 2. Random Effects in Model 1

Groups	Name	Variance (logit)
Excerpt	Intercept	2.92
Participant	Intercept (Avante-garde)	0.09
	Blues	0.17
	Classical	0.07
	Country	0.26
	Folk	0.05
	Jazz	0.12
	Pop/Rock	0.20
	Rap	0.13
	R&B	0.17
	Reggae	0.27
	Religious	0.24
	Theater	0.24

though we can't be 95% confident that the best estimate for Reggae and Rap differ, we can be 80% confident that they do. Still, even the 80% confidence intervals for many of the genre estimates overlap to a large degree, so no reasonable claim can be made that these genres differ in intelligibility (compare Blues and R&B for instance). This is due to the large variability in intelligibility observed within most of the genres. Within-genre variation is accounted for in the model by the random intercept estimates for excerpts, which are represented by gray dots in Figure 1. As can be seen, all the sampled genres (except Classical) have excerpts with estimated success rates above 90%. Particularly variable are Pop/Rock and Avante-garde, each with both highly intelligible and unintelligible excerpts. As example, the Pop/Rock stimuli include "Death Metal" excerpts which received intelligibility scores of zero (no participant was able to identify a single word) alongside "Pop" excerpts that achieved scores close to 100%. Within-genre variance is no doubt reflective of the anticipated imperfection in any broad genre categorization. In summation, though the significant log-likelihood test suggests that the variability in intelligibility across the 248 excerpts is significantly reduced by the binning of excerpts into the twelve broad genre categories, the proportion of variance accounted for is modest. Thus, musical features that influence intelligibility are only partly associated with specific genres, and can vary widely independently of genre.

Random effect estimates for Model 1 are given in Table 2. The values given in the table are estimates of the amount which the fixed effect slope estimates (Table 1) randomly vary in the population (i.e., beyond the current sample) of excerpts and participants. Thus, the model attributes a relatively large variance (2.93 in logit scale) in intelligibility to random variation between excerpts which is not accounted for by genre or excerpt length. This is the within-genre variance that has been discussed, and plotted, in Figure 1. The model estimates of random variance due to participants is relatively modest, with the largest value of 0.26 representing variation in participant's ability to understand Country music specifically. Compare this to the estimate for Folk music, with a relatively small value of 0.05 indicating that participants are expected to be relatively consistent in their abilities to decipher lyrics in Folk music.

Questionnaire items. We next sought to compare participants' questionnaire responses to their success rates. For the "attentiveness" and "importance" questionnaire items, participants' responses were reasonably well distributed through the entire seven-point scale available to them. Mean responses for each scale were 3.5 for attentiveness ($SD = 1.7$) and 4.5 for importance ($SD = 1.6$).

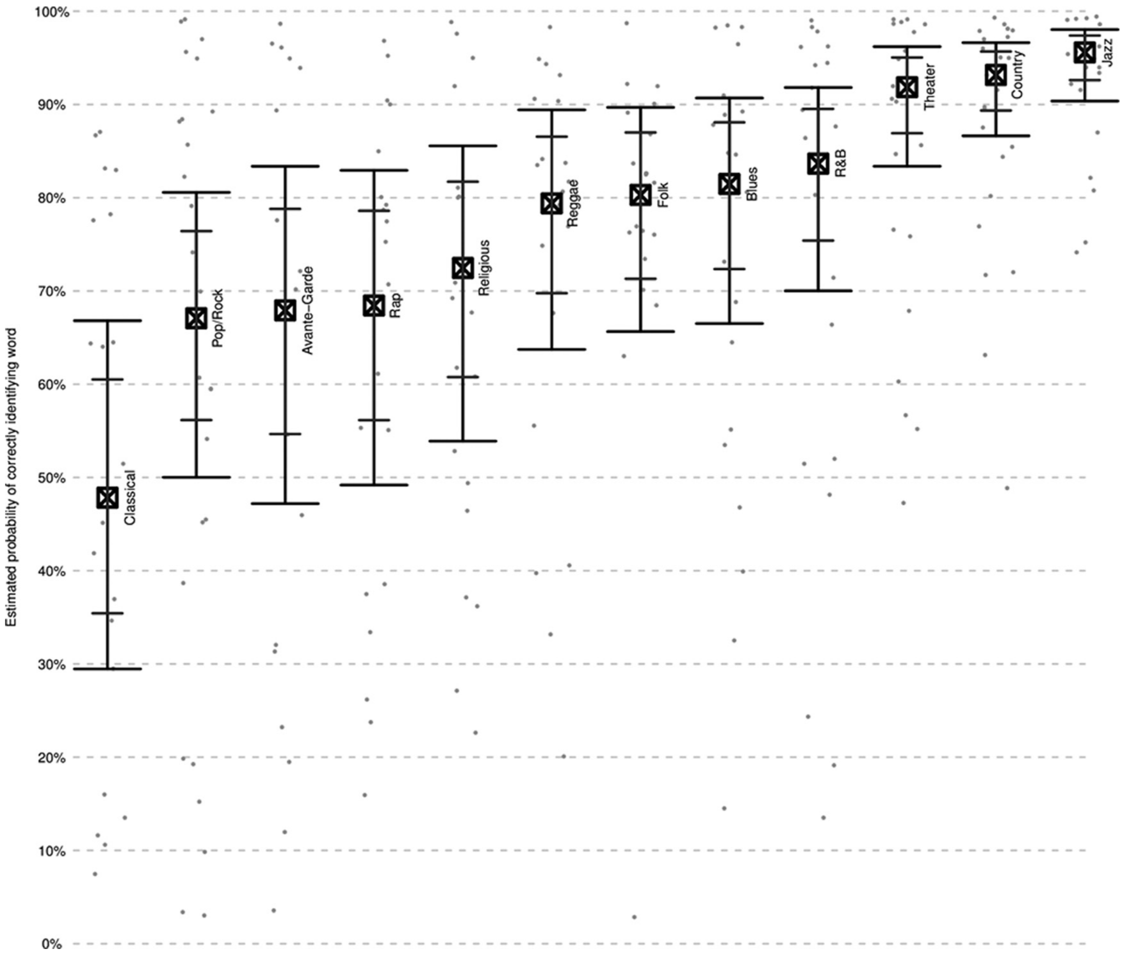

FIGURE 1. Intelligibility estimates for twelve genres

Regarding singing experience, no participant identified themselves as a "Professional Singer" but the remaining five values (1–5 on the six-point scale) were distributed reasonably as well ($M = 3.0$, $SD = 1.25$). Correlations between the three rankings were not large enough to suggest redundancy. Genre familiarity ratings were also well distributed across genres and participants ($M = 3.9$, $SD = 2.02$). The genres exhibited notable differences in their mean familiarity as presented in Table 3, with Pop/Rock in particular clearly the most familiar genre. It is notable that these familiarity rankings do not correlate well with genre intelligibility estimates, suggesting that genre familiarity does not play a strong role intelligibility.

Between participants, the observed proportion of correct responses ranged from a low of 60.2% to a high of 77.1%. A new model was constructed, identical to Model 1 except adding the questionnaire responses as fixed effects. Log-likelihood tests for each questionnaire item found only the importance item to be significant (Table 4). Specifically, an approximately 9% increase in the odds of correctly identifying a word is predicted for every one unit increase on the importance scale. This correlation gives no hint as to causality: it may be that listeners who consider lyrics to be important are better at deciphering sung lyrics, or that those listeners who are better able to decipher the lyrics are more likely to consider the lyrics important. Note, however, that no

TABLE 3. Mean Genre-familiarity Ratings by Genre

Genre	Mean	*SD*
Avante-garde	2.1	1.6
Religious	2.9	2.2
Reggae	3.1	1.7
Blues	3.5	1.7
Folk	3.7	1.9
R&B	4.1	1.5
Classical	4.2	2.2
Jazz	4.3	1.8

TABLE 4. Questionnaire-item Fixed Effects

Factor	χ^2	*df*	*p*	*Slope (logit)*	*SE*	*z*	*p*
Importance item	4.60	1	.03	0.08	0.04	2.25	.02
Attentiveness item	0.86	1	.35	−0.03	0.04	−0.87	.38
Singing experience item	0.76	1	.38	0.05	0.05	0.87	.38
Genre familiarity item	0.77	1	.38	0.02	0.02	0.89	.37

correction for multiple tests has been employed in this analysis. A more conservative approach might consider even the importance item to be nonsignificant.

Spontaneous listening task. Finally, recall that participants were given a spontaneous lyric transcription task at the beginning of the experiment. This task was intended as an implicit measure of listeners' attentiveness to lyrics, to complement the explicit measures in the questionnaire. Of the 25 participants who took part in the trial, nine did not spontaneously attend to the lyrics at all. Of the 16 participants who did attend, nine correctly identified all the words. Since the spontaneous nature of the task allowed only one response per participant, power for statistical tests is necessarily limited. Correlations between spontaneous responses and participants' recognition rates and questionnaire responses were not significant. A χ^2 test comparing those who attended at all to those who did not was also underpowered and not significant. These underpowered statistics notwithstanding, at face value it appears that roughly a third of participants fail to attend to lyrics in a casual listening situation. Ultimately, both implicit and explicit measures of the variation in listener attentiveness to lyrics will require further study.

DISCUSSION

The results broadly match our expectations. It is slightly surprising that the average word-recognition rate when listening to real musical excerpts was roughly consistent with the recognition rate of 73.7% found by Collister and Huron (2008) using isolated words as stimuli. Taken alone, this would seem to imply that actual musical and lyrical context has little effect on intelligibility. However, as expected we did find significant differences between genres (and more so between excerpts), with some genres/excerpts much more intelligible and some much less intelligible than the average. This does suggest that different elements of musical and meaningful context may either degrade or improve intelligibility. Results for measures of interpersonal variation in lyric deciphering abilities and tendencies were mixed, with most questionnaire items not being significant predictors. Though no strong claims about interpersonal variation can be made at this time, the results do suggest that this line of questioning may be a fertile topic for future research.

During the coding process a possible problem became apparent. In one example, the validated lyric "He stole the most, even the crown" was misheard as "Killin' the moose with our feet on the ground." The participant scored two successes (for the two instances of "the") and five failures for this excerpt, a recognition rate of 28%. However, it might be argued that even this low score exaggerates the participant's success, as none of the meaningful content of the lyric was correctly deciphered. In a contrasting example, mishearings of words like "the" or "a" often dragged down the scores of otherwise successful hearings. For example, in the response mentioned above to the lyric beginning "a short plaid skirt" the participant's score was lowered for leaving out the "a." In this case, it might be argued that the important meaningful content of the lyric is being understood, yet the scoring system is taking away points based on unimportant function words. These problems suggest that the current approach to scoring participants' responses might be improved. Since a principal advantage of this study over previous research is using real lyrics in context, simply scoring words in isolation independent of their meaningful context may be inappropriate. These considerations led us to a post hoc reformulation of our original question: rather than "what percentage of words do listeners understand?" one might ask "what percentage of the meaning are listeners understanding?" To explore the effects these issues might have on the scores, two additional recodings were undertaken. In one recoding only content words were scored; function words like "a" and "the" were ignored. In a second less systematic recoding, minor mistakes involving things like tense, gender, perspective, or either semantically or functionally similar words were awarded partial scores. For instance, "a" in place of "the," "he" in place "she," or "there" in place of the contraction "there's" received partial scores rather than zero. In another example, the phrase "my brightest hopes" misheard as "my greatest hopes" received a partial score due to the essentially equivalent meaning. Overall, if the misheard lyric seemed to reflect the meaning of the actual lyric this "lenient" score was relatively high. A randomly selected one third of excerpts were recoded using these methods. The "content only" recoding had no appreciable effect on the participants' scores: as in the examples involving "the" above, the "content only" recoding sometimes improved scores but in other cases lowered scores, leading to no overall change. The "lenient" recoding did boost scores slightly but not sufficiently to alter any overall conclusions (a global average of 73% rather than 72%). Hence, it seems that the original scoring scheme is not easily improved; consequently we elected not to recode the entire data set.

An important consideration during the planning phase of the experiment was the appropriate duration of the musical excerpts. Recall that variability in excerpt length between genres was considered as a potential confound to our experimental design, as longer excerpts may be more difficult to remember and write down. A statistically significant relationship might suggest that the observed effects are attributable to limitations of short-term memory rather than acoustic factors of intelligibility. Accordingly, two measures of excerpt length (log-seconds and number of words) were included in all regression models. No statistically significant correlation was found between the number of words in an excerpt and its recognition rate. As clear examples, two of the wordiest excerpts (with 11 and 14 words respectively) received near perfect scores, whereas many excerpts with few words scored poorly. However as reported above, length in seconds did appear to be a significant predictor suggesting that longer excerpts may have suffered from memory limitations. Interpretation of the regression coefficient given for log-seconds in Table 1 is not intuitively obvious. Roughly, the estimate implies that approximately tripling the excerpt length in seconds halves the odds of success (Figure 2). The worst scoring genres, Classical, Avante-garde, and Pop/Rock did also have the longest average excerpt length. However, the best scoring genre, Jazz, had the fourth longest excerpts, averaging less than a second shorter then Classical excerpts. In addition, the genre with by far the shortest average excerpt length, Rap (5.1 seconds average), is below average scoring in intelligibility. Examining Figure 2 suggests that the observed effect may be largely driven by a few extremely long excerpts (over 20 seconds) with poor intelligibility scores. However, limiting the analysis to excerpts shorter than ten seconds still reveals a significant effect, though attenuated. In any case, the estimates in all reported models have already taken into account the influence of excerpt length. Thus, there is no reason to believe that the length of excerpts is responsible for the intelligibility differences observed between genres. An obvious question which awaits future research is how syllable rate per second (not considered here) influences intelligibility.

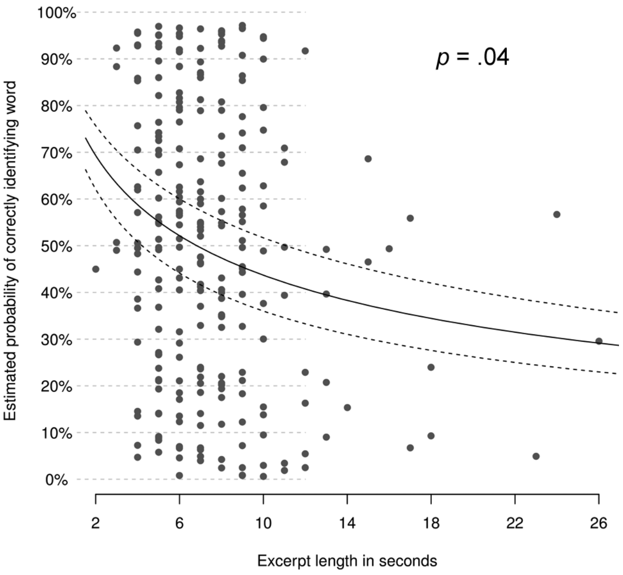

FIGURE 2. Excerpt intelligibility estimates by excerpt length in seconds

POST HOC ANALYSES

Establishing that there are intelligibility differences between genres is rather straightforward and unsurprising. Further research is necessary in order to establish what the musical and meaningful contextual elements are that improve or degrade intelligibility. As a start to this research, two simple post hoc tests were devised: one test exploring the role of pitch height in intelligibility and a second exploring the relative loudness of the voice.

Pitch height. To explore the influence of pitch height in the current data, the melodies of each excerpt were transcribed and the mean pitch per excerpt was calculated. Singer gender was also coded. Twenty-three excerpts featuring unpitched (spoken or screamed) vocals were excluded from this analysis, as were 14 excerpts which featured both male and female singers. A new mixed-effects regression model was created, similar to Model 1, with mean pitch, singer gender, and their interaction as fixed effects, as well as genre and excerpt length in log-seconds as covariate fixed effects. Random intercepts for excerpts and participants, as well as random slopes for genres within participants were included in the model. Log-likelihood comparison tests found both excerpt pitch height and singer gender to be nonsignificant (Table 5). Gender and interaction slopes in Table 5 are given as ± since they represent the differences between

TABLE 5. Fixed Effects in Pitch-height Model

Factor	χ^2	*df*	*p*	*Slope (logit)*	*SE*	*z*	*p*
Gender	0.4	1	.53	±4.46	3.48	1.28	.20
Mean pitch	1.6	1	.21	−0.08	0.04	−1.77	.08
Pitch X Gender	1.5	1	.23	±0.07	0.06	1.22	.22

the two levels of the gender category, and would be positive or negative depending on which group is considered the reference level. For log-likelihood tests gender and mean pitch were compared to a model without the interaction term. These results may seem to be contrary to previous research. However, previous research has suggested that pitch height has a negative effects on intelligibility only when the pitch is extremely high in the soprano vocal range. The highest mean excerpt pitch in the current sample is G5 with the majority of the excerpts far lower in pitch, averaging around D4 for female singers. Thus, none of the current excerpts approach the range at which pitch height is expected to have an impact on intelligibility.

Relative loudness of voice. In a second post hoc analysis, we sought to explore the influence of "audio mix" on intelligibility, specifically the relative loudness of the voice compared to the instrumental accompaniment. It seems highly likely that masking or other interference from instruments could be a major detriment to intelligibility. Objective measurements of relative intensity, sound pressure level differences, spectrum, or masking effects in complex musical textures are beyond the scope of this paper. As a simple subjective measure, the first author listened to each excerpt and rated on a six-point scale how clearly the vocal part could be heard over the instrumentation. On this scale, a score of one represents a mix where the voice is significantly masked by the instruments, whereas a five represents a mix where the voice is clearly and easily audible above the instrumentation. Values of six were reserved for excerpts which were solo (unaccompanied) voice. As in the previous analyses, subjective relative loudness ratings were added to Model 1 as fixed effects. A log-likelihood comparison test found the relative loudness ratings to be significant predictors of success rates ($x^2 = 20.3, df = 1, p < .01$). The relationship between the relative loudness rating and excerpt intelligibility is illustrated in Figure 3. The significant upward slope corresponds to a 58% increase in the odds of correctly identifying a word for every one unit increase in the volume rating. Thus, as predicted, a louder voice compared to the accompaniment does appear to lead to better intelligibility.

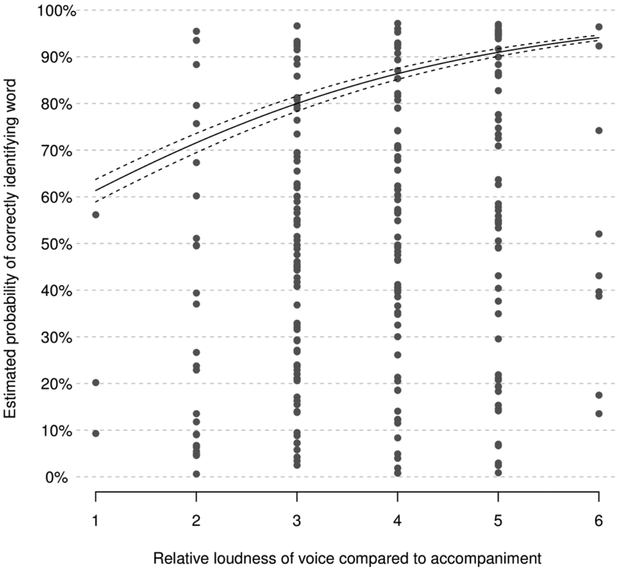

FIGURE 3. Excerpt intelligibility estimates by loudness ratings

Results from these post hoc analyses should be interpreted with caution, as the excerpts were not prepared with these tests in mind. Further obvious factors that likely influence intelligibility include syllable rate, the amount of reverberation, and the frequency content (timbre) of the accompaniment. Finally, aspects of singing style may prove to have the greatest effect on intelligibility. Understanding the influence of these and other musical (rhythmic setting) and poetic (word use) factors on intelligibility will require further research.

Experiment 2

The first experiment allowed participants only a single hearing of each excerpt. However, in everyday life listeners typically listen to songs numerous times. An obvious question is how listeners' hearing of lyrics improves with multiple hearings. In general, it seems likely that listeners will improve their transcriptions with multiple hearings. However, there are many anecdotal accounts where an initial mishearing of the lyrics is simply duplicated and persists with successive listenings. Accordingly, a second experiment was conducted to test the effect of multiple hearings. An interesting observation from post-experiment interviews during the first experiment was the high degree of confidence that participants had in their interpretations. Thus, during the second experiment a measure of confidence was also included in the experimental design.

TABLE 6. Fixed Effects in Model 2

Factor	χ^2	*df*	*p*	*Slope (logit)*	*SE*	*z*	*p*
Length in log-seconds	1.52	1	.28	−0.89	0.70	−1.26	.20
Listening trial	29.0	4	< .01				
1st vs. 2nd-5th trial				0.60	0.08	7.10	< .01
2nd vs. 3rd-5th trial				0.16	0.06	2.58	< .01
3 rd vs. 4th-5th trial				0.08	0.06	1.33	.18
4th vs. 5th trial				0.05	0.07	0.77	.44

Method

STIMULI

A subset of the excerpts from Experiment 1 were used as stimuli. To avoid ceiling effects, excerpts that averaged a recognition rate of greater than 80% in the first experiment were discarded, leaving 125 excerpts. To focus power on both average-scoring (70% range) and poorly scoring examples, a stratified sample was gathered, with ten excerpts randomly selected from excerpts that averaged between 70% and 80% correct in the first experiment, and thirty excerpts randomly sampled from excerpts that averaged less than 70%. Each participant heard a random sample of twenty excerpts from these forty.

PARTICIPANTS

A different group of 16 participants was recruited from the same pool.

PROCEDURE

The method for Experiment 2 was identical to Experiment 1 in most respects. The computer interface was altered so that each excerpt was heard a total of five times. After each hearing the participant typed in their interpretation of the lyrics as before. When they pressed ENTER, the program moved on to the next hearing. During repeated hearings the typed text was removed from the screen. This was done to encourage participants to rethink and improve their interpretations as much as possible after each hearing. In addition, each time participants typed a response they were asked to indicate on a seven-point scale how confident they were that their current interpretation of the lyrics was the correct one. The end-points of the scale were labeled “not at all confident” and “totally confident.” Participants were allowed to leave the field blank if their interpretation had not changed since their previous hearing, but they were required to indicate their confidence after each hearing. After the fifth hearing, they were given special warning that they would now hear a new excerpt.

Analyses

CODING

Responses were scored as in the first experiment.

RESULTS

Success rates. The results for the first hearings of each excerpt were consistent with the results from the first experiment: In the first experiment, the mean proportion of correct responses to the forty excerpt subset was 45% while in the second experiment the first listening trail average was 47%. A new mixed-effects logistic regression model was constructed. The main fixed effect of interest was listening trial. Treating listening trail as a continuous variable did reveal a significant effect. However, visual inspection of success rates across trials suggested that this apparent linear trend is created by improvements after the first two listenings, after which success rates did not increase. Thus, listening trial was recoded as a categorical variable using Helmhert coding. Helmhert coding compares each category step by step to the mean of the remaining categories. Log-seconds was included as a covariate but genre was not, because the excerpts sampled for the seconds study were not well spread across the genres and were in fact largely taken from genres which did not appear to be significantly different in the first experiment. Random intercepts for participant and excerpt, as well as slopes for listening trial within both, were included in the model. Details of the model results are presented in Table 6 and Figure 4. As can be seen, on second and third hearings participants were significantly more likely to correctly identify words compared to earlier hearings, though the improvement on the third listening is less than a quarter of the second. Beyond the third listening, success rates appear to plateau.

Confidence ratings. Recall that participants were asked to indicate their confidence in the correctness of their answer after each listening on a seven-point scale. Participants made good use of the entire seven-point scale available to them. Participant’s confidence significantly

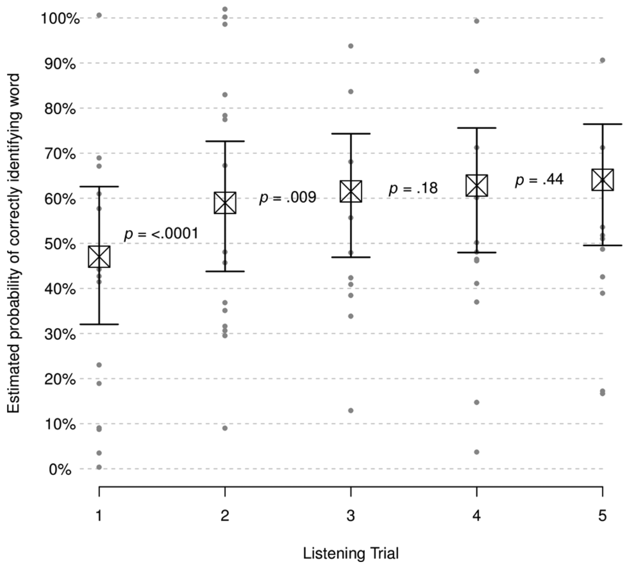

FIGURE 4. Excerpt intelligibility estimates over five hearings

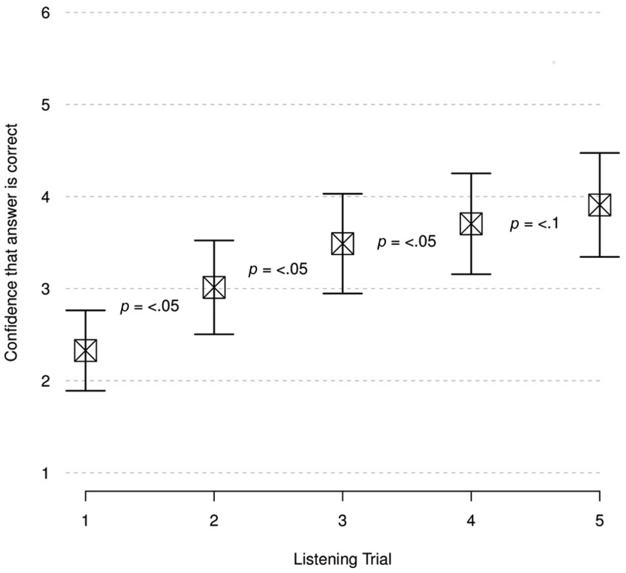

FIGURE 5. Participant confidence ratings over five hearings

TABLE 7. Fixed Effects in Confidence Model

Factor	χ^2	*df*	*P*	*Slope*	*SE*	*t*
Proportion correct	423	1	< .01	3.11	0.14	22.63
Listening trial	23	4	< .01			
1st vs. 2nd-5th trial				0.88	0.13	6.65
2nd vs. 3rd-5th trial				0.60	0.13	4.50
3 rd vs. 4th-5th trial				0.27	0.11	2.62
4th vs. 5th trial				0.17	0.09	1.93

predicted their success rates ($\chi^2 = 557$, $df = 1$, $p < .01$) with every one unit increase in confidence predicting a 50% increase in the odds of correctly identifying a word. However, participants' confidence also significantly increased over listening trials, and upon inspection of the data it appeared likely that this increase was greater than what would be warranted by actual improvement in success rates. A new model was constructed which was identical to Model 2, except that the model predicts confidence rather than success rate (log-seconds were not significant in predicting confidence and are not included in this model). As can be seen in Table 7 and Figure 5, participants' confidence continued to increase even after the third trial, possibly even on the fifth trial (the *t* value of 1.93 is marginal). This is despite the fact that, as we've established, actual success rates flattened out after the third listening. Thus, it appears that participants' confidence that they have identified the lyrics correctly continues to increase upon repeated hearing, even when there are mistakes in their interpretation. On the whole, the results of Experiment 2 suggest that listeners' interpretations of lyrics stabilize after an average of three hearings, beyond which they are no longer likely to improve their transcriptions. Rather, it seems that listeners simply become increasingly confident that their interpretation is correct, whether or not it actually is.

Conclusion

In agreement with earlier studies, the experiments reported here provide further empirical evidence indicating difficulties in deciphering sung lyrics. Unlike other studies, this study shows that significant difficulties arise even in more ecologically valid listening situations spanning a wide variety of vocal musical styles. Moreover, these difficulties persist even with repeated listenings, and even when listeners focus on the task of catching the lyrics. Singers, composers, and lyricists cannot take for granted that their words will be intelligible, even by attentive listeners. Our results suggest that various musical attributes, including singing style, vocal tessitura, syllable-rate, instrumentation, and audio mix influence intelligibility.

In addition, the current research suggests that further study of individual differences across listeners may be warranted. Recall that listener attitude towards the importance of the lyrics was positively correlated with intelligibility scores. It may be that listeners who consider lyrics to be important are more motivated to decipher sung lyrics, or that those listeners who are better able to decipher the lyrics are more likely to consider the lyrics important.

Future research may explore both individual differences as well as the specific influences of various musical attributes, and how these attributes may be manipulated so as to improve intelligibility.

Author Note

Correspondence concerning this article should be addressed to David Huron, School of Music, 1866 College Road, Ohio State University, Columbus, OH 43210. E-mail: huron.1@osu.edu; or Nathaniel Condit-Schultz, School of Music, 1866 College Road, Ohio State University, Columbus, OH 43210. E-mail: natsguitar@gmail.com

References

Benolken, M. S., & Swanson, C. E. (1990). The effect of pitch-related changes on the perception of sung vowels. *Journal of the Acoustical Society of America*, *87*, 1781-1785.

Burleson, R. (1992). Functional-relationships of language and music — The 2-profile view of text disposition. *Linguistique*, *28*(2), 49-63.

Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: A reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1235-1249.

Collister, L. B., & Huron, D. (2008). Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review*, *3*, 109-125.

Coren, S., & Hakstian, A. R. (1992). The development and cross-validation of a self-report inventory to assess pure-tone threshold hearing sensitivity. *Journal of Speech and Hearing Research*, *35*, 921-928.

Gilchrist, A. L., Cowan, N., & Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult aging: Recall of fewer, but not smaller chunks in older adults. *Memory, 16*, 773-787.

Hollien, H., Mendes-Schwartz, A. P., & Nielsen, K. (2000). Perceptual confusions of high-pitched sung vowels. *Journal of Voice*, *14*, 287-298.

Johnson, R., Huron, D., & Collister, L. (2012). Music and lyrics interactions and their influence on recognition of sung words: An investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming. *Empirical Musicology Review*, *9*(1), 2-20.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

Smith, L. A., & Scott, B. L. (1980). Increasing the intelligibility of sung vowels. *Journal of the Acoustical Society of America*, *67*, 1795-1797.