DETECTING FOREIGN ACCENTS IN SONG

Pre-published version

Final version: Phonetica 76(6), 2019

Marly Mageau

Canadian Transportation Agency, Regulatory Affairs

15 rue Eddy, Gatineau, QC J8X 4B3, Canada

Can Mekik

Department of Cognitive Science, Rensselaer Polytechnic Institute

Troy, NY 12180, USA

Ashley Sokalski

Institute of Cognitive Science, Carleton University

1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada

Ida Toivonen

Ida.Toivonen@carleton.ca

Institute of Cognitive Science and School of Linguistics & Language Studies, Carleton University

1125 Colonel By Drive, Ottawa, ON

K1S 5B6, Canada

1 Introduction

Professional singers often shift their accents when they sing. For example, many British, New Zealand and Australian singers shift their accents in song so that they are easily mistaken for Americans. Trudgill (1983); Simpson (1999); Morrissey (2008); Beal (2009); and Gibson (2010) argue that these singers target a particular accent (consciously or unconsciously) for sociological reasons: the singers strive to sing with an accent that is associated with a specific genre. In pop, rock and country music, the target accent is often American.

Previous studies focus on experienced or trained singers, and mostly singers in genres broadly associated with particular accents. We investigate whether singing *inherently* masks certain markers of accent. If it does, then accent should also be more difficult to detect in inexperienced singers singing a song associated with a genre that neither the singer nor the listener associates with a specific accent. To our knowledge, our study is the first to investigate the accents of untrained and inexperienced singers. We specifically aimed to eliminate the influence of training, practice and genre-specific accents as potentially confounding factors. By "training", we do not mean exclusively classical vocal training, which many professional singers in fact lack. Instead, we intend to draw a distinction between singers who have extensive experience singing, possibly accompanied by vocal coaching and explicit training, and people who have no experience of organized singing practice or training.

The studies cited above focus on professional singers who are native English speakers shifting away from the accent associated with their native dialect when they sing. The experiments presented in the present paper are more similar to the study of Hagen et al (2011), who examine the accents of non-native English speakers in song. Hagen et al. recruited eleven Dutch choir singers with English as a second language (L2) and six native English-speaking choir singers. These subjects were recorded reading and singing passages of familiar songs. The researchers manipulated one sentence of the reading recordings in three different ways. In the first, the durations of the non-native speakers were matched with the duration of a native English speaker at the segmental level. In the second, pitch (f0) was monotonized. In the third, the duration was manipulated like in the

first and the pitch was also monotonized. These manipulations were intended to explore the role pitch and duration have in identifying an accent. Twenty native English listeners were asked to listen to and judge the stimuli on native accent authenticity using a 7-point scale, with '1' meaning 'very strong foreign accent' and '7' meaning 'native'.

Hagen et al (2011) found that the listeners rated non-native speakers as having less of a perceived accent in song compared to speech. They also found that listeners gave higher ratings (closer to native) when listening to the manipulated recordings. The researchers concluded that f0 and duration are important accent markers, but accents cannot be signalled with f0 and duration in song, since the melody imposes its own f0 and duration on the text.

Our first study is similar to Hagen et al.'s with respect to the general research questions and underlying ideas, but the design differs significantly. Our study has a larger sample size, and the stimuli involve recordings of native speakers of a number of different languages, in contrast to Hagen et al.'s study, which focuses exclusively on Dutch. Furthermore, unlike Hagen et al., we do not make use of manipulated stimuli.

The experiments presented here concern the perception of accent in song and speech (Study 1), but also the production of pitch, duration, and quality of vowels in these two modalities (Study 2 and 3). Song clearly affects pitch, and Gibson (2010) points out that the rhythm of the melody affects duration as well. There are some further studies on professional singers that suggest that the vowel formants are also somewhat influenced, although the effects are not as drastic: The first formant (F1) is higher in song than in speech, and the second formant (F2) is lower in front vowels in song than in speech (Sundberg, 1969; Sundberg and Skoog, 1997, 1999; Clermont, 2002). However, these studies note that F1 and F2 differences of speech and song vary across individuals, singing techniques, genres and pitch. Since all of these studies were performed on professional singers, the results might be due specifically to singers being trained to manipulate their vocal tract while singing.

Vowel duration plays an important role for intonation and stress, but is also linguistically relevant in other ways. English does not have phonemically contrastive vowel length; that is, vowel duration alone does not signal a difference in meaning. However, vowel duration is an important cue for other distinctions, such as vowel height

(Heffner, 1937; Lehiste, 1970; Toivonen et al, 2014) and consonant voicing (House and Fairbanks, 1953; Peterson and Lehiste, 1960, a.o.). Vowels are shorter before voiceless than voiced consonants (House and Fairbanks, 1953, a.o.). This generalization holds cross-linguistically, but to different extents in different languages. For example, the effect has been reported to be generally stronger in English than in French, even though vowels in French are also shorter before voiceless consonants (Chen, 1970; Mack, 1982). In addition, previous studies have shown variation in duration differences due to voicing according to a variety of factors other than target language. de Jong (2004) found larger vowel duration differences according to consonant voicing in stressed than in unstressed contexts. Mack (1982) reports that bilingual French-English speakers have overall different duration ratios due to voicing of the following consonants than monolingual speakers. In a study including only bilingual French-English speakers, Muldner et al (2019) found greater duration differences due to consonant voicing in code-switched contexts than in monolingual contexts. One of the questions addressed in our third study is whether singing affects the contrast in vowel duration before voiced and voiceless consonants.

All three studies presented in this paper concern the accents of native and non-native English speakers. Most previous studies comparing native and non-native accents have naturally focused on accents in speech. Research has shown that a wide range of linguistic factors are relevant when listeners judge foreign accents in speech. For example, Magen (1998) found that English-speaking listeners judging Spanish-accented English were sensitive to multiple features, including syllable structure, final [s] deletion, and consonant manner. Gubbins and Idemaru (2011); Chan et al (2017); and others have shown that vowel quality plays a crucial role in the perception of foreign accents. Studies focusing on production have also found significant differences between native and non-native speakers in vowel formants (Munro, 1993; Baker and Trofimovich, 2005; Oh et al, 2011; Yang and Fox, 2017, a.o.). A number of studies have further shown that vowel duration is a significant marker of a foreign accent (Jonasson and McAllister, 1972; Munro, 1993; Tajima et al, 1997). Flege (1993) specifically found that when non-native English speakers produced a large difference in vowel duration cued by the voicing of the following consonant, their perceived foreign accent was lowered. Drawing on these previous results, we were interested in whether some markers of accentedness would be affected by

the modality of song.

In Study 1, native speakers of English were asked to guess whether the voices on recordings belonged to native or non-native speakers, and they were then asked to rate the accentedness of the recordings they had classified as non-native. Study 2 performed acoustic analyses on the recordings used for the perception judgements in study 1. The recordings were samples of native and non-native speakers of English singing and reading passages of the song *Twinkle, Twinkle Little Star*. The fundamental frequency (pitch), duration, and first and second formants of vowels were measured and compared. Study 3 was a repetition of Study 2, with the difference that the recordings targeted a more limited sample of sung and spoken words. Study 3 measured vowel pitch, duration and quality like Study 2, and it also compared vowel duration before voiced and voiceless consonants in song and in speech. Even though vowels are typically longer before voiced than before voiceless consonants, the extent to which this generalization holds depends on a number of factors as outlined above. We set out to test whether the effect of voicing on duration is stronger in speech than in song, and whether there are any differences between native and non-native speakers in this regard.

2 Study 1

The goal of this study was to investigate whether English-speakers are better at detecting accents in speech than in song. English speakers were asked to judge whether sung and read passages were recorded by native or non-native speakers. They were further asked to rate the accents of the speakers/singers that they had guessed to be non-native speakers. The basic design of the study follows Mageau et al (2015). Based on the results reported in Mageau et al (2015) and also in Trudgill (1983); Gibson (2010); and Hagen et al (2011) (described in section 1), we hypothesized that the participants would find it easier to detect foreign accents in speech than in song.

2.1 Materials

Twelve female speakers (ages 20-44) were recruited from the Ottawa area to be recorded. Six were native speakers of Canadian English and six were non-native speakers of En-

glish. We recognize that "native speaker" is a complex concept, but for the purposes of this paper, we consider a person who has been immersed in English since their early childhood as a native speaker. The native speakers were all exposed to English from birth and had little exposure to any other language. The non-native speakers had not been immersed in an English-dominant environment until adulthood, but they all had some training in English growing up. The non-native speakers' native languages were French (two speakers), Tamil, Farsi, Romanian, and Spanish.[1] We did not administer tests to gauge language proficiency. The fact that the non-native speakers first languages and English language proficiency varied are not confounding variables as this project does not address specific questions about language transfer or L1 influence. None of the twelve speakers were trained in music or singing.

The materials consisted of the first two verses of *Twinkle Twinkle Little Star* (see Appendix I) and the first paragraph of *Goldilocks and the Three Bears* (see Appendix 2). Neither *Twinkle Twinkle Little Star* nor *Goldilocks and the Three Bears* is inherently associated with any particular regional accent of English, as they are sung or read to children throughout the English-speaking world. We recorded the speakers reading the *Goldilocks* passage, and we also recorded them both reading and singing *Twinkle Twinkle*. They sang accompanied by a recording of a piano playing just the melody of the song, without the harmonies. We used four recordings per speaker: one reading of Goldilocks, one reading of *Twinkle Twinkle*, and one singing of *Twinkle Twinkle* that was used twice: one with the accompaniment and one without the accompaniment.[2]

The recordings were organized into four blocks so that each block included twelve passages: three readings of Goldilocks recordings, three readings of *Twinkle Twinkle* passages, three *Twinkle Twinkle* songs with accompaniment, and three *Twinkle Twinkle* songs without accompaniment. Each block contained only one recording per speaker. For example, if Block A included a recording of Speaker A reading *Goldilocks*, then no other recordings of Speaker A would be included in Block A. Each block included recordings of all twelve speakers/singers (six native and six non-native speakers of English). The

[1] We did not select these particular speakers for any specific reasons, we simply recorded the participants who volunteered.

[2] The speakers/singers listened to the accompaniment in headphones as their song was recorded, so it was easy to include or exclude the channel with the accompaniment.

recordings within each block were randomized.

A Tascam HD-P2 portable solid-state recorder paired with both head and table mounted Audio Technica microphones (one microphone for each recording channel) was used to record all participant sessions.

2.2 Methods

We recruited forty native speakers of Canadian English as participants. The participants were recruited through internet advertisement and flyers that were posted at Carleton University. Participants were guaranteed complete anonymity and no biographical information was collected beyond the fact that they were all self-reported speakers of Canadian English. They each received a small gift certificate for their participation. The participants did the study individually, not in a group. Each participant was asked to listen to one of the four blocks of recordings described in *Materials*. Ten participants listened to each block, and the recordings within the block were randomized for each participant. After every recorded passage, the participants were asked to guess whether the recorded voice belonged to a native or a non-native speaker of English. Each time they guessed that the speaker/singer was a non-native speaker, the participants were further asked to rate the accent of the people on those recordings on a 9-point scale, where a 1 corresponded to "closer to native Canadian English" and a 9 corresponded to "less close to native Canadian English". We used the wording "close to native" and "less close to native" because that wording is commonly used in studies addressing the perception of accents (see, e.g. McCullough 2013). We specified *Canadian* English because all of our native speakers were speakers of Canadian English, and we thought it would be possible that potential similarity between the accent of one or more of the non-native speakers and some other native variety of English could affect the results.

2.3 Results

We collected 480 guesses in total (40 participants, 12 guesses, fully crossed design). Of these guesses there were 70 incorrect guesses (15%) and 410 (85%) correct guesses. Table 1 displays the full distribution of incorrect guesses.

Table 1: Incorrect responses by condition

Condition	Incorrect guesses	Percent
Singing with accompaniment	32	46%
Singing without accompaniment	17	24%
Reading *Twinkle, Twinkle*	14	20%
Reading *Goldilocks*	7	10%

Table 1 illustrates that most of the incorrect responses are in the two singing conditions (70%). This suggests that a listener who has difficulty detecting a foreign accent has greater difficulty when the speakers are singing than when they are reading. Table 1 also illustrates that the participants guessed the accent incorrectly especially often when the recorded speakers sang with musical accompaniment (46%). Participants guessed correctly most often when listening to the prose reading passage, *Goldilocks and the three bears*.

A proportion test was used in order to determine whether the distinction between song and speech indicated by the descriptive statistics was significant. We tested our results against the null hypothesis (according to the null hypothesis, there would be no significant difference in number of incorrect guesses between the song and speech conditions). The proportion test showed that there was a statistically significant increase of incorrect responses for the singing conditions compared to the reading conditions ($\chi^2 = 11.2$, $p < 0.01$). These results are consistent with the hypothesis that it is more difficult to detect a foreign accent in song than in speech.

When participants guessed that a speaker/singer was a non-native speaker of English, they were asked to rate how close to a native speaker the person on the recording sounded (1 = close to Canadian English to 9 = less close to Canadian English). Table 2 shows the average ratings for the different conditions.

Table 2: Listeners judgements of accents

Condition	Ratings, *M (SD)*
Singing with accompaniment	4.5 (2.5)
Singing without accompaniment	5.1 (2.4)
Reading *Twinkle, Twinkle*	6.1 (2.3)
Reading *Goldilocks*	6.2 (2.5)

Table 2 illustrates that native English listeners rate speakers as having less of an accent when they sing than when they speak. Within the singing condition, the ratings were lower for songs accompanied by piano than for the a cappella song. Within the reading condition, the ratings were lower for the verse passage (*Twinkle, Twinkle*) than for the prose passage (*Goldilocks*).

We performed a linear mixed effects analysis to illustrate the relationship between rating and speech or song. The modality variable (speech versus song) was entered into the model as a fixed effect.

Participants, speakers and conditions were entered as random effects, as these variables could play a role in the rating. P-values were obtained using a likelihood ratio test of the full model with and without the interested effect (i.e., the difference in speech and song). The likelihood ratio test found that there was a significant difference between the two models which indicates that speech and song have an effect on ratings ($\chi^2(1)= 5.39$, $p= 0.02$), where the rating in song is lower by about $1.1 +/-0.4$ (standard error).

In sum, the results from the guessing task together with the results from the rating task indicate that accents are less noticeable in song than in speech.

3 Study 2

Study 2 makes use of the materials collected for Study 1 to compare the acoustic properties of vowels in song and in speech. Specifically, we compare the sung samples of *Twinkle, Twinkle* to the read samples of *Twinkle, Twinkle*. Our main goal was to investigate acoustic differences between song and speech in native and non-native speakers. We hypothesized that the melody and rhythm of the music will influence the duration and

pitch of vowels. Since previous studies have shown that F1 is overall higher and F2 of front vowels is lower in song than in speech, we also hypothesized that we might see a difference in formants between the singing and the speaking condition.

Munro (1993) and others (see Section 1) have also found differences between native and non-native speakers in the quantity and quality of vowels, and a secondary goal was therefore to explore whether there were systematic differences between the groups of speakers (native and non-native) in our sample.

3.1 Methods

This study investigates the pitch, duration and quality of spoken and sung vowels in native and non-native speakers. The vowel measurements were conducted using PRAAT (Boersma and Weenink, 2016). Because different vowels have intrinsic properties, we limited the number of vowels to /ɪ/ and /ʌ/. These specific vowels were chosen because they occur more often in the verses than other vowels: /ɪ/ occurs 16 times and /ʌ/ occurs 10 times (see Appendix 1 and 2). We include one sung and one spoken passage from 12 speakers, 24 passages in total. The study thus includes measurements from 624 vowels. Vowel duration was measured from the offset of the pre-vocalic consonant to the onset of the post-vocalic consonant. The frequencies of fundamental frequency (f0) and the first two formants (F1-F2) were measured at the midpoint of the vowel, as is common in acoustic analyses of vowels (e.g., Flege et al 1997).

Since the measured vowels occur in different phonetic environments and were uttered by different speakers with different linguistic backgrounds and in different modalities (song and speech), the data are not normally distributed. Furthermore, there tends to be more variance in song than in speech (Gibson, 2010, 75). The assumptions for parametric tests are therefore not met, so we used pairwise Wilcoxon tests for our analyses, following Gibson (2010) who investigated comparable data sets. We first aggregated the data from each speaker and condition.[3] That is, we calculated the average duration, f0, F1, and F2 values for vowel (/ɪ/ and /ʌ/) and condition (speech and song). We then performed analyses for each vowel (/ɪ/ and /ʌ/) separately for native and non-native speakers. We used Wilcoxon signed-rank tests for these comparisons. We also performed analyses to

[3]We aggregated data for all analyses except pitch range (see section 3.2.2).

compare the duration, f0, F1, and F2 values of native and non-native speaker participants, and for these between-speaker analyses, we used Wilcoxon rank sum (Wilcoxon-Mann-Whitney) tests.

3.2 Results

3.2.1 Duration

The average durations for the vowels /ɪ/ and /ʌ/ for native speaker participants and non-native speaker participants are presented in Table 3.

Table 3: Duration (msec), *M (SD)*

	Speech		Song	
	/ɪ/	/ʌ/	/ɪ/	/ʌ/
Native	66 (27)	92 (36)	153 (91)	172 (59)
Non-native	74 (25)	104 (33)	185 (79)	195 (62)

Wilcoxon signed-rank tests were conducted in order to determine the statistical significance of the durational differences in song compared to speech. The native speaker participants produced longer vowels in song than in speech, both for the /ɪ/ vowel (V = 21, p = 0.03) and the /ʌ/ vowel (V = 21, p = 0.03). The non-native speakers also produced longer vowels in song than in speech, both for the /ɪ/ vowel (V = 21, p = 0.03) and the /ʌ/ vowel (V = 21, p = 0.03).[4]

Wilcoxon rank sum tests further showed that non-native and native speaker participants' durations were not significantly different in the speaking condition (/ɪ/: W = 29, p=0.09; /ʌ/: W = 24, p=0.39), and also not for the /ʌ/ vowel in the singing condition (W = 30, p=0.06). However, the native speaker /ɪ/ vowels were shorter than the non-native speaker /ɪ/ vowels in the singing condition (W = 5, p = 0.04).

In sum, both the native and non-native participants produced longer vowels in song than in speech. The native and non-native speakers differed significantly with respect to

[4]The fact that the Wilcoxon statistics are identical here is explained by the fact that all values on one list (the values from the song condition) were higher than the paired values on the other list (from the speech condition).

vowel duration only when producing /ɪ/ vowels and only when singing.

3.2.2 Pitch

Our participants produced higher pitch (as measured by fundamental frequency, f0) in song than in speech. Higher pitch in song has also been reported in previous studies (Sundberg and Skoog, 1999). The average f0 values for the vowels /ɪ/ and /ʌ/ are given in Table 4.

Table 4: f0 (Hz) of /ɪ/ and /ʌ/, *M (SD)*

	Speech		Song	
	/ɪ/	/ʌ/	/ɪ/	/ʌ/
Native	188 (23)	169 (29)	247 (65)	229 (56)
Non-native	218 (51)	191 (47)	311 (66)	278 (55)

Pitch was significantly higher for both the native speaker participants (/ɪ/ vowel: V = 21, p = 0.03; /ʌ/ vowel: V = 21, p = 0.03) and the non-native speaker participants (/ɪ/ vowel: V = 21, p < 0.03; /ʌ/ vowel: V = 21, p = 0.03).

The pitch of the native speakers was also compared to the pitch of the non-native speaker participants. There was no significant difference in the speech condition (/ɪ/ vowel: V = 29, p = 0.09; /ʌ/ vowel: V = 24, p < 0.39). In the song condition, the /ʌ/ vowels did not differ (W = 30, p = 0.06), but the /ɪ/ vowels were significantly higher for the non-native speaker participants (W = 31, p = 0.04).

The pitch ranges for each speaker were also examined. The ranges were calculated by subtracting the lowest f0 for each vowel from the highest f0 for each vowel. This gave two values per speaker (one for each vowel). The average pitch ranges for speakers are given in Table 5.

Table 5: f0 ranges (Hz), *M (SD)*

	Speech		Song	
	/ɪ/	/ʌ/	/ɪ/	/ʌ/
Native	71 (42)	74 (51)	118 (39)	115 (41)
Non-native	98 (42)	70 (44)	186 (26)	147 (58)

The pitch range was significantly greater in song than in speech both for native speaker participants (V = 67, p = 0.03) and non-native speaker participants (V = 76, p < 0.01). There was no significant difference in range between native and non-native speakers in the speech condition (W = 81, p = 0.62), but the difference was significant in the song condition (W = 116, p = 0.01).

3.2.3 Vowel height (F1)

Recall that previous researchers have found that speakers produce overall higher F1 (corresponding to vowel lowering) in song than in speech (Sundberg 1969, and others). However, the participants of this study did not display any shift in vowel height (neither lowering nor raising) when singing. The average F1 values are given in Table 6.

Table 6: F1 values (Hz) of the vowels /ɪ/ and /ʌ/, *M (SD)*

	Speech		Song	
	/ɪ/	/ʌ/	/ɪ/	/ʌ/
Native	526 (79)	705 (137)	521 (116)	705 (149)
Non-native	412 (104)	664 (153)	444 (115)	707 (140)

Wilcoxon signed rank tests did not yield significant results for either vowel in native speakers (/ɪ/ vowel: V = 11, p = 1; /ʌ/ vowel: V = 20, p = 0.82) or non-native speakers (/ɪ/ vowel: V = 17, p = 0.22; /ʌ/ vowel: V = 14, p = 0.56).

Wilcoxon rank sum tests did not show any significant differences between native and non-native speaker participants. There was no difference in the speech condition (/ɪ/ vowel: W = 30, p = 0.06; /ʌ/ vowel: W = 20, p = 0.82), or in the song condition (/ɪ/ vowel: W = 29, p = 0.09; /ʌ/ vowel: W = 15, p = 0.70).

3.2.4 Vowel backness (F2)

Previous studies have found that singing correlates with a shift in vowel backness, as measured by F2 (Sundberg 1969 and others). Specifically, front vowels have been shown to display lower F2 in singing than in reading, indicating backing of front vowels in song. In contrast to the previous studies, our data did not indicate any consistent F2 formant

shift when comparing speech and song. The average F2 values are given in Table 7.

Table 7: F2 values (Hz) of the vowels /ɪ/ and /ʌ/, *M (SD)*

	Speech		Song	
	/ɪ/	/ʌ/	/ɪ/	/ʌ/
Native	1963 (325)	1503 (149)	2115 (464)	1572 (199)
Non-native	2008 (600)	1340 (193)	1978 (695)	1444 (288)

There was no significant difference between song and speech in F2 for native speaker participants (/ɪ/ vowel: V = 3, p = 0.16; /ʌ/ vowel: V = 1, p = 0.06) or non-native speaker participants (/ɪ/ vowel: V = 27, p = 0.18; /ʌ/ vowel: V = 18, p = 0.16).

Wilcoxon rank sum tests did not show any significant differences between native and non-native speakers participants. There was no difference in the speech condition (/ɪ/ vowel: W = 27, p = 0.18; /ʌ/ vowel: W = 6, p = 0.06), or in the song condition (/ɪ/ vowel: W = 12, p = 0.39; /ʌ/ vowel: W = 12, p = 0.39).

3.3 Summary of Study 2

Study 1 showed that English speakers did better at identifying native speakers of English when listening to spoken samples than when listening to sung samples. In Study 2, we analyzed the vowels in the recordings used as stimuli for Study 1 in order to gauge whether any of the core vowel characteristics shifted significantly so that native and non-native vowels were more similar in the sung data. We did not find evidence for any such pattern. Instead, we found that both native and non-native speakers shifted their pitch and duration when singing, whereas the F1 and F2 stayed stable across conditions, again for both native and non-native speakers.

4 Study 3

Our third study had two goals. The first was to replicate Study 2 with the same basic design but more constrained stimuli: instead of the lyrics of Twinkle Twinkle, the participants repeated a limited set of words in speech and song. The second goal was to ex-

plore further how music influences the natural durational patterns of speech in untrained singers. We measured duration, f0, F1, and F2 as we did in Study 2. We hypothesized that the results pertaining to duration and pitch would be the same as in Study 2, but we expected that the more limited stimuli could reveal potential differences in vowel quality (F1 and F2) that were not evident in the previous study. We also investigated whether the difference in duration before voiced and voiceless consonants (see section 1 above) is maintained in song, and whether native and non-native speakers pattern the same in this regard. Flege (1993) found that the distinction in vowel length cued by consonant voicing is a marker of native English accent, and other studies have found that various factors can affect this distinction (Mack 1982; de Jong 2004; Muldner et al 2019; see section 1). These observations together with the finding that accents are more difficult to detect in song than in speech led us to hypothesize that the distinction in vowel duration cued by consonant voicing might be lesser in song than in speech.

4.1 Methods

Fourteen speakers were recruited to participate in the study.[5] Recruitment took place via the Carleton University Institute of Cognitive Science SONA system. The speakers were all female and between 18 and 24 years old. The native speakers had been exposed to English since birth and had little exposure to a second language. The non-native speakers had not been immersed in an English-speaking environment until adulthood. However, all speakers had some English exposure and/or training while growing up. Seven of the speakers were native speakers of English and seven had learned English as a second language. The second language speakers' first languages were different varieties of Arabic (four speakers) and Chinese (three speakers: two Mandarin and one Cantonese). Participation was voluntary, and participants received compensation in the form of a 1% grade increase in a first or second-year Cognitive Science course, which they were enrolled in at the time of their participation in the study. None of the speakers were trained singers.

A Tascam HD-P2 portable solid-state recorder paired with both head and table mounted Audio Technica microphones (one microphone for each recording channel) was used to

[5] All participants of the previous studies were guaranteed anonymity and we did not have their contact details. We therefore could not approach the speakers/singers of studies 1 and 2 for this study.

record all participant sessions. A metronome set at a comfortable pace (120 beats per minute) was used in order to ensure that the participants maintained an even pace when they read and sang the target words. The acoustic analysis software PRAAT allowed for segmentation and vowel length analysis.

We recorded each participant individually. In the speech portion of the study, speakers were asked to read the words *bus, buzz, bet* and *bed* 42 times each with the help of a metronome set to 120 beats per minute for 30 seconds to keep their pace steady. The order in which they read the words was randomized. Each word was said 42 times because that number corresponds to the number of syllables in one verse of *Twinkle Twinkle Little Star*, and we wanted to ensure that the two conditions were as similar to each other as possible.

In the singing portion of the study, speakers were asked to sing each of the words *bus, buzz, bet* and *bed* to the tune of the first verse of *Twinkle Twinkle Little Star*. They sang one verse at a time, each time repeating the same word. Each participant therefore sang four verses in total, one for each of the four words. For example, a participant would first sing repetitions of the word *bus* to the melody of *Twinkle Twinkle*, and then repetitions of the word *buzz*, etc. The order in which they sang the words was randomized. A metronome was again set to 120 beats per minute so that speakers would maintain a steady pace while singing.

The condition order was alternated so that half of the speakers began their recording session with the speech portion of the study and half with the singing portion of the study.

The sound files were then imported into PRAAT so that vowel length could be extracted for both the speech and song conditions. Because the last syllable of each line of *Twinkle Twinkle Little Star* is a half note instead of the usual quarter note, the last word for each line of the song condition was removed, as it would have an artificially long vowel.[6] This means we removed six tokens of each word. We also removed words that had errors in them, and we finally removed random words to make sure we had the same number of tokens in the speech condition and the song condition. 36 tokens of each of the four words remained per speaker and condition. We thus analyzed 144 spoken and 144 sung words per participant, 4032 words in total.

[6]Note that no half notes were included in Study 2. They were automatically excluded as a result of the two specific vowels that were investigated.

4.2 Results

The stimuli were more restricted in this study than in Study 2, but the F1 and F2 values nevertheless failed to display normal distribution. We therefore analyzed F1 and F2 with Wilcoxon tests. However, we did not find evidence that the pitch and duration displayed non-normal distribution, so we analyzed those measurements with ANOVAs.

4.2.1 Duration

The native speakers' spoken vowels averaged 164 msec and their sung vowels 200 msec. The non-native speakers also have longer vowels in song with their spoken vowels averaging 150 msec and their sung vowels 201 msec. The average measurements for vowel duration for all four words, calculated separately for song and speech and native and non-native speakers, are provided in Table 8.

Table 8: Vowel duration for words ending in a voiced *(buzz, bed)* or voiceless *(bus, bet)* consonant (msec), *M (SD)*

	Speech			
	buzz	*bus*	*bed*	*bet*
Native	207 (53)	144 (53)	176 (53)	130 (52)
Non-native	165 (51)	137 (51)	165 (51)	133 (52)
	Song			
	buzz	*bus*	*bed*	*bet*
Native	238 (53)	191 (52)	205 (52)	165 (52)
Non-native	213 (52)	193 (51)	213 (51)	185 (51)

To determine if vowel duration was lengthened in the singing condition dependent on speaker type (native and non-native), a mixed ANOVA was used with speaker type as the between-participant variable, and modality (speech and song), voicing, and vowel type (/ɛ/ and /ʌ/) as within-participant variables.

The results showed that there was no significant main effect of speaker type ($F(1, 12) = 0.247$, $p = .063$). These results are consistent with the study reported in section 3.2.1 above: there was no significant overall difference in vowel duration between our native

and non-native speaker participants in either study. However, there were significant main effects of modality, vowel type, and voicing.

The results for modality ($F(1, 12) = 19.07$, $p < 0.001$) showed that vowels are longer in the singing condition (M = 202, SD = 40) than in the speaking condition (M = 156, SD = 50). This result replicates the findings of previous studies and the findings of the study reported in 3.2.1.

The results for vowel type ($F(1, 12) = 5.51$, $p = 0.03$) showed that /ʌ/ vowels (M = 181, SD = 52) are longer than /ɛ/ vowels (M = 172, SD = 49) for both native and non-native speakers.

In addition, the results for voicing ($F(1, 12) = 47.67$, $p < 0.001$) showed that vowels are significantly longer before voiced consonants (M = 196, SD = 50) than before voiceless consonants (mean = 157, SD = 40). This confirms that both sets of speakers maintained the voiced-voiceless distinction in both the speech and singing conditions. The aspects of rhythm and intonation dictated by the melody in the singing condition do not fully neutralize this basic phonological pattern of English.

There was a near significant interaction between speaker type and voicing ($F(1, 12) = 4.08$, $p = 0.06$). The non-significant (at the 0.05 level) trend towards an interaction prompted further analysis of the data. The descriptive statistics suggested that the trend is due to the voicing of the consonant having a particularly large effect on vowel duration in the native speaker participants' speech condition. The average ratio of vowel duration (duration before voiced consonants to duration before voiceless consonants) in native speakers in the speech condition was 1.40:1 (SD = 0.23), which is similar to what has previously been reported in the literature for English lax vowels (Peterson and Lehiste, 1960; Raphael, 1972, a.o.). In the song condition, the average ratio was only 1.25:1 (SD = 0.18), and a paired Welch t-test (two-tailed) comparing the two indicated that the difference was significant, $t(13) = 2.60$, $p = 0.02$. The ratio was also greater in the speech than the song conditions for non-native speakers. The non-native participants' average ratio was 1.23:1 (SD = 0.21) in the speech condition and 1.13:1 (SD = 0.17) in the song condition. A paired Welch t-test (two-tailed) showed that the difference was significant $t(13) = 2.47$, $p = 0.03$. Comparisons between native and non-native speakers revealed that the ratio difference was significant in the speech condition but not the song condition.

The two-tailed Welch t-test results for the ratio comparison between native and non-native speakers in speech was t(25.7) = 2.10, p = 0.046, and in song t(25.9) = 1.76, p = 0.09. Together, these results suggest that the consonant voicing has a smaller effect on vowel duration in song than in speech.

4.2.2 Pitch

Pitch was also analyzed with a mixed ANOVA. Speaker type was the between-participant variable, and the within-participant variables were modality (speech and song), voicing, and vowel type (/ɛ/ and /ʌ/). Several of the participants used breathy voice when singing and/or creaky voice when speaking, which meant that many of their f0 values could not be read. For this reason, two native and two non-native speakers were excluded from the pitch analysis.

The main result concerning pitch in Study 2 was higher pitch in song than in speech, and this was replicated here: there was a significant effect of modality (F(1, 10) = 44.11, $p < 0.001$) with higher pitch in the singing condition (M = 258, SD = 63) than in the speaking condition (M = 188, SD = 66). With the exception of the /ɪ/ vowel in the song condition, Study 2 found no differences between native and non-native speakers. The results of this study similarly showed that there was no significant main effect of speaker type (F(1, 10) = 0.402, p = .54). There were also no main effects of voicing (F(1, 10) = 1.80, p = .22) or vowel type (F(1, 10) = 1.44, p = .26), and there were no significant interactions.

4.2.3 Vowel height (F1)

Several of our speakers (both native and non-native speakers) pronounced their *buzz* vowel and especially their *bus* vowel quite far back, which meant that the F1 and F2 values often overlapped and could not reliably be extracted. To avoid including potentially incorrect data in our analysis, we therefore excluded the *bus, buzz* measurements from our analyses of F1 and F2.

Study 2 did not find any evidence that the participants shifted their vowel height when singing, and this study yielded the same result. Wilcoxon signed rank tests comparing the F1 of sung and spoken vowels did not yield significance in native speakers (V = 41, p =

0.50; speech M = 664, SD = 148; song M = 633, SD = 146) or non-native speakers (V = 60, p = 0.67; speech M = 681, SD = 94; song M = 693, SD = 114). Wilcoxon rank sum tests also did not show any significant differences between native and non-native speaker participants. There was no difference in the speech condition (W = 95, p = 0.91), or in the song condition (W = 124, p = 0.25).

4.2.4 Vowel backness (F2)

Study 2 did not find any evidence for a shift in vowel backness in song, and we similarly failed to find evidence for a shift in F2 in non-native speakers in this study. However, this study did find that F2 values were significantly lower in song than in speech in native speakers. Wilcoxon signed rank tests comparing the F2 of sung and spoken vowels yielded significance in native speakers (V = 10, $p < 0.01$; speech M = 1595, SD = 437; song M = 1328, SD = 363), but not in non-native speakers (V = 47, p = 0.76; speech M = 1562, SD = 281; song M = 1503, SD = 259).

Study 2 did not find any evidence of significant differences between native and non-native speakers in song and speech, and neither did this study. Wilcoxon rank sum tests showed no difference in the speech condition (W = 82.5, p = 0.49), or in the song condition (W = 126.5, p = 0.2).

4.3 Summary of Study 3

The results of this study (Study 3) were very similar to the results of Study 2. The *duration* finding that vowels are longer in song than in speech for both native and non-native speakers was replicated. The duration results further indicate that vowels are longer before voiced than voiceless consonants in both song and speech, and this was also found for both native and non-native speakers. An exploration of a near-significant trend further revealed that the difference in vowel duration caused by the voicing of the following consonant seems to be greater in speech than in song, especially for native speakers.

Study 3 also replicated the finding that *pitch* is higher in song than in speech, both for native and non-native speakers. Like Study 2, Study 3 found no significant shift in F1 in song. Unlike Study 2 but in line with previous results reported by Sundberg (1969) and

others, the native speakers in our study did produce vowels with lower F2 in song than in speech. However, for non-native speakers the results were the same as in Study 2: there was no evidence of a shift in F2 in non-native speakers.

5 Discussion

The goal of this project was to investigate the perception and production of foreign accent markers in song. Our first main research question was whether it is more difficult to detect accents when listening to someone who is singing than when listening to someone who is speaking. Results from previous studies have suggested that it is indeed more difficult to detect accents in speech than in song (Trudgill, 1983; Simpson, 1999; Hagen et al, 2011, a.o.) and our study confirmed those claims and findings.

Our study was novel in that it specifically targeted speakers with no musical training. Previous studies have focused on professional solo singers from different genres (Trudgill, 1983; Simpson, 1999; Gibson, 2010), or experienced choir singers (Hagen et al, 2011). We recorded and studied samples from people who are not experienced singers nor musically trained in other ways. We played the recordings to native English speaker participants, who did significantly better at guessing whether the people on the recordings were native speakers when they were reading, compared to when they were singing. The participants further ranked the accents as less native-like when they listened to a read passage than when they listened to a sung passage.

Our second main research question was how singing affects the acoustic characteristics of sounds. We specifically focused on vowels. We investigated pitch (f0), duration, vowel height (F1) and vowel backness (F2). We found that vowels were produced with a higher pitch (fundamental frequency) and a greater pitch range in song than in speech. These findings are consistent with previous research, and they are not surprising, as singing requires following the melody of the song. We also found that vowels are generally longer in song than in speech. Vowels are the main carriers of pitch in each syllable, and we suggest, following Gibson (2010), that the relative long duration in song is due to the rhythmic nature of song: vowels need to be lengthened in order to follow the beat.

Previous research has also found certain shifts in F1 in song, but our results did not

corroborate those findings. Some previous research has found evidence of lower F2 in front vowels, which we also found, but only in native speakers and only in one of the two production studies (Study 3). We hypothesize that our studies largely failed to replicate the F1 and F2 findings from the literature because we only included untrained singers. Previous studies have focused on trained singers, and part of singing practice is to (consciously or unconsciously) learn how to modify the production of sounds to make them carry the tones better, to be more audible, and indeed to sound more true to the accent they are targeting (see, e.g., Marshall 1953; LaBouff 2008; Wall et al 2009; Christiner and Reiterer 2015). It is therefore likely that the consistent shifts in vowel pronunciation (evident in F1 and F2 measurements) that previous studies have found are due to the fact that experienced singers learn to modify their sounds, and especially their vowels, as they sing. Untrained singers like the ones we recorded have not learned to consistently shift their pronunciation when they sing.

Study 3 compared the duration of vowels in words ending in voiced and voiceless consonants, and we found that the lengthening effect of consonant voicing on preceding vowels remains in song. A trending interaction led us to further explore the effect of voicing. Our data indicated that that the difference in vowel duration due to consonant voicing is greater in speech than in song for both native and non-native speakers. However, there was an interesting difference between song and speech: the ratios for native and non-native speakers were significantly different in speech, but not in song. In other words, the native and non-native speakers are more similar in song than in speech with respect to the effect of consonant voicing on vowel duration. We consider this a potentially interesting topic of future research, because if this finding can be replicated, it could mean that singing can lessen certain language-specific phonological effects. This in turn might be one of the reasons why native and non-native speakers sound more alike when they sing than when they speak.

It could also be interesting to further explore the recognition of accents in different types of song and speech. Although we did not specifically explore this question, the descriptive statistics seem to suggest that the presence of accompaniment made accents even harder to detect. The accompaniment consisted simply of the melody played on piano, and the accompaniment on the recording was not loud. Nevertheless, previous

studies show that noise renders the detection of accents more challenging (Munro, 1998; Adank et al, 2009; Gordon-Salanta and Yeni-Komshian, 2010), so it may be the case that any extra sounds can distract from possible accent markers. It would also be interesting to explore whether accents are more readily detected in prose (such as *Goldilocks*) than in verse (such as *Twinkle, Twinkle*). Since verse forms superimpose a given rhythm (even when it is read, not sung), perhaps this metrical rhythm serves to partially mask markers of accents that have to do with intonation and stress, as signalled by duration, pitch and loudness.

In this paper, we focus on reporting the overall significant effects evident in the data. However, visual inspection reveals that there is variation between and within individuals, especially among the non-native speakers and especially with respect to F1 and F2 values. This is expected, as the speakers have different linguistic backgrounds and the non-native speakers have different levels of fluency in English. A possible topic for future study could be to investigate this variation more carefully. For example, perhaps there are different degrees of within-speaker variation in song than in speech. In general, studies comparing native speakers to non-native speakers who all have the same first language would allow for more systematic investigation of specific aspects of accents in speech and song.

Hagen et al (2011) suggest that it is more difficult to detect accents in song than in speech because the rhythm and melody of the song partially masks intonation and stress. The results from the studies presented here are consistent with their proposal: our studies consistently showed significant shifts in duration and pitch (acoustic realizations of rhythm and melody) between song and speech, while F1 and F2 (acoustic characteristics not directly tied to rhythm and melody) mostly remained stable across the two conditions (with the exception of F2 values in native speakers in Study 3). The rhythm and melody of a song overrides some of the natural intonational patterns of a speaker. Accent markers that relate to stress and intonation therefore do not come across as clearly in song as in speech. The results of this study are thus compatible with the results of the studies reported in Anderson-Hsieh et al (1992); Boula de Mareuil and Vieru-Dimilescu (2006); Ulbrich and Mennen (2016); and Silva and Barbosa (2017) that conclude that prosody has a crucial effect on the perception of accented speech.

Acknowledgements

We would like to thank Catherine Boucher, Ryan Grenon, Raj Singh, and Emily Wang for feedback and assistance at various stages of this project. We also thank James Kirby for pointing us to an important article. Finally, we are very grateful to the editor Sonia Frota and two anonymous reviewers for many important questions and suggestions.

Statement of ethics

The study protocols for our studies have been approved by the Carleton University Research Ethics Board A (CUREB-A), files #102248 and #104691. All participants have given their written informed consent.

Author Contributions

The authors are listed purely alphabetically. Mageau and Sokalski contributed to the design, data collection, analysis and writing of this paper. Toivonen contributed to the design, analysis and writing, and Mekik contributed to all aspects in the initial stages of the project.

Appendix 1

The lyrics for the first two verses of *Twinkle Twinkle Little Star*.

Twinkle, twinkle, little star
How I wonder what you are
Up above the world so high
Like a diamond in the sky
Twinkle, twinkle little star
How I wonder what you are

When the blazing sun is gone
When he nothing shines upon

Then you show your little light
Twinkle, twinkle, all the night
Twinkle, twinkle, little star
How I wonder what you are

Appendix 2

The beginning passage of *Goldilocks and the Two Bears*

Once upon a time there were three bears: A father bear, a mother bear and a little bear. They lived all together in a yellow house in the middle of a big forest. One day, Mother Bear prepared a big pot of delicious hot porridge for breakfast. It was too hot to eat, so the bears decided to go for a walk while waiting for the porridge to cool. Near the forest lived a little girl named Goldilocks.

References

Adank P, Evans B, Stuart-Smith J, Smith SK (2009): Comprehension of familiar and unfamiliar native accents under adverse listening conditions. Journal of Experimental Psychology: Human Perception and Performance 35:520–529.

Anderson-Hsieh J, Johnson R, Koehler K (1992): The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. Language Learning 42(4):529–555.

Baker W, Trofimovich P (2005): Interaction of native- and second-language vowel system(s) in early and late bilinguals. Language and Speech 48(1):1–27.

Beal JC (2009): You're not from New York City, you're from Rotherham: Dialect and identity in Bristish indie music. Journal of English Linguistics 37(3):223–240.

Boersma P, Weenink D (2016): Praat: Doing phonetics by computer (version 6.0.18) [computer program], downloaded from http://www.praat.org/.

Boula de Mareuil P, Vieru-Dimilescu B (2006): The contribution of prosody to the perception of foreign accent. Phonetica 63:247–267.

Chan KY, Hall MD, Assgari AA (2017): The role of vowel formant frequencies and duration in the perception of foreign accent. Journal of Cognitive Psychology 29(1):23–34.

Chen M (1970): Vowel length variation as a function of the voicing of the consonant environment. Phonetica 22:129–159.

Christiner M, Reiterer SM (2015): A Mozart is not a Pavarotti: Singers outperform instrumentalists on foreign accent imitation. Frontiers in Human Neuroscience 9(8):482–490.

Clermont F (2002): Systemic comparison of spoken and sung vowels in formant-frequency space. In: Proceedings of the 9th Australian SST, Hong Kong, pp 124–129.

de Jong K (2004): Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. Journal of Phonetics 32(4):493–516.

Flege JE (1993): Production and perception of a novel, second-language phonetic contrast. Journal of the Acoustical Society of Americ 93:1589–1608.

Flege JE, Bohn OS, Jang S (1997): Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics 125(4):437–470.

Gibson A (2010): Production and perception of vowels in New Zealand popular music. Master's thesis, Auckland University of Technology.

Gordon-Salanta S, Yeni-Komshian GH (2010): Recognition of accented English in quiet and noise by younger and older listeners. Journal of the Acoustical Society of America 128(5):3152–3160.

Gubbins L, Idemaru K (2011): Foreign accent production and perception: An acoustic analysis of non-native Japanese. Journal of the Acoustical Society of America 130(4):24–28.

Hagen M, Kerkhoff J, Gussenhoven C (2011): Singing your accent away, and why it works. In: Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, pp 799–802.

Heffner RMS (1937): Notes on the length of vowels. American Speech 12(2):128–134.

House A, Fairbanks G (1953): The influence of consonant environment upon the secondary acoustical characteristics of vowels. Journal of the Acoustical Society of America 25:268–277.

Jonasson J, McAllister R (1972): Foreign accent and timing: an instrumental study. PILUS 14:11–40.

LaBouff K (2008): Singing and Communicating in English: A Singer's Guide to English Diction. Oxford University Press, Oxford.

Lehiste I (1970): Suprasegmentals. The MIT Press, Cambridge, MA.

Mack M (1982): Voicing-dependent vowel duration in English and French: monolingual and bilingual production. Journal of the Acoustical Society of America 71:173–178.

Mageau M, Mekik C, Wang E (2015): Foreign accents in speech and song. In: The Ottawa Conference for Undergraduate and Masters Students Conference, Ottawa.

Magen HS (1998): The perception of foreign-accented speech. Journal of Phonetics 26:381–400.

Marshall M (1953): The Singer's Manual of English Diction. Schirmir Books, New York.

McCullough EA (2013): Acoustic correlates of perceived foreign accent in non-native English. PhD thesis, The Ohio State University.

Morrissey F (2008): Liverpool to Louisiana in one lyrical line: Style choice in British rock, pop and and folk singing. In: Locher M, Strässler J (eds) Standards and Norms in the English Language: Contributions to the sociology of language, De Gruyter, Berlin.

Muldner K, Hoiting L, Sanger L, Blumenfeld L, Toivonen I (2019): The phonetics of code-switched vowels. International Journal of Bilingualism 23(1):37–52.

Munro M (1998): The effects of noise on the intelligibility of foreign-accented speech. Studies in Second Language Acquisition 20:139–154.

Munro MJ (1993): Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. Language and Speech 36:39–66.

Oh GE, Guion-Anderson S, Aoyama K, Flege JE, Akahane-Yamada R, Yamada T (2011): A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting. Journal of Phonetics 39(2):156–167.

Peterson GE, Lehiste I (1960): Duration of syllable nuclei in English. Journal of the Acoustical Society of America 32(6):693–703.

Raphael LJ (1972): Preceding vowel duration as a cue to the perceptual separation of cognate sounds in American English. Journal of the Acoustical Society of America 51:1296–1303.

Silva CC, Barbosa PA (2017): The contribution of prosody to foreign accent: A study of Spanish as a foreign language. Loquens 4(2):1–14.

Simpson P (1999): Language, culture and identity: With (another) look at accents in pop and rock singing. Multilingua 4(18):343–367.

Sundberg J (1969): Articulatory differences between spoken and sung vowels in singers. Speech Transmission Laboratory Quarterly Progress and Status Report (STLQPSR) 10(1):33–46.

Sundberg J, Skoog J (1997): Dependence of jaw opening on pitch and vowel in singers. Journal of Voice 11(3):301–306.

Sundberg J, Skoog J (1999): Formant frequencies in country singers' speech and singing. Journal of Voice 13(2):161–167.

Tajima K, Port R, Dalby J (1997): Effects of temporal correction on intelligibility of foreign-accented English. Journal of Phonetics 25(1):1–24.

Toivonen I, Blumenfeld L, Gormley A, Hoiting L, Logan J, Ramlakhan N, Stone A (2014): Vowel height and duration. In: Steindl U, Borer T, Fang H, García-Pardo A, Guekguezian P, Hsu B, O'Hara C, Ouyang IC (eds) Proceedings of WCCFL 32, Cascadilla Proceeding Project, Somerville, MA, pp 64–71.

Trudgill P (1983): On dialect: Social and geographical perspectives. Blackwell, Oxford.

Ulbrich C, Mennen I (2016): When prosody kicks in: The intricate interplay between segments and prosody in perceptions of foreign accents. International Journal of Bilingualism 20(5):522–549.

Wall J, Caldwell R, Allen S, Gavilanes (2009): Diction for Singers, 2nd Edition. Pst, Dallas.

Yang J, Fox RA (2017): L1-l2 interactions of vowel systems in young bilingual Mandarin-English children. Journal of Phonetics 65:60–76.