Construction and validation of the Seattle Singing Accuracy Protocol (SSAP): An automated online measure of singing accuracy

Peter Q. Pfordresher, University at Buffalo, State University of New York

Steven M. Demorest, Northwestern University

The Seattle Singing Accuracy Protocol (SSAP) arose from a workshop held in October 2013 (Demorest et al., 2015). A small group of researchers who study accuracy of pitch matching – a critical component of the larger range of skills involved in singing – determined a minimal set of tasks necessary to measure pitch accuracy in a way that could be useful for a broad range of research contexts. Thus, in comparison to other measures like the AIRS Test Battery of Singing Skills (ATBSS, Cohen, 2015) and the Sung Performance Battery (SPB, Berkowska & Dalla Bella, 2013), the SSAP was intended to be both brief and highly specific in its focus. In addition, the original intent was that the SSAP include a fully automated measure of singing accuracy based on the acoustic signal. This feature was intended to make the process of analysis easier for researchers who do not typically focus on acoustic measures of accuracy. It also allows members of the general public to use the SSAP and to obtain immediate summary statistics relating to their own pitch accuracy.

A conceptual overview of the tasks included in the SSAP and their rationale was reported in a special issue of the journal Music Perception that arose from the workshop (Demorest et al., 2015). We provide an updated overview here, given that more details are now available concerning the interface. However, the main purpose

CITATION: Pfordresher, P. Q., & Demorest, S. M. (2020). Construction and validation of the Seattle Singing Accuracy Protocol (SSAP): An automated online measure of singing accuracy. In F. Russo, B. Ilari, & A. Cohen (Eds). Routledge Companion to Interdisciplinary Studies in Singing: Vol. 1 Development (pp. 322-333). London: Routledge.

of this chapter is to offer a detailed account of the automated procedures in the SSAP, and to report on results of the SSAP following the collection of a large number of cases to date. We will use these tests to further consider the validity of the SSAP's automated scoring. These analyses also provide the opportunity to reflect on the effectiveness of the SSAP in its present form, and to consider planned modifications to improve its functioning in the near future.

Overview of the SSAP

The SSAP (Demorest & Pfordresher, 2015) is designed to measure how accurately one can reproduce musical pitch by singing, both in the context of matching pitch in an imitative "call and response" framework, and while singing songs from memory. The protocol also measures important possible correlates of singing accuracy, including simple pitch discrimination and questionnaire responses concerning age, gender, musical background, musical self-concept and other demographic information. The SSAP is organized into seven subtests described below; a menu bar is shown on the side throughout the test showing the participant where they are in the procedure.

At the start of the protocol, participants are given the choice of clicking *General Public* or *Study Participant*. The latter choice is followed by a prompt to enter their study ID number. No personal identifiers are ever collected. All participants are then asked to indicate their informed consent with the procedure. Then the program tests the sound output and microphone input for the computer to make sure that recording is functioning properly. The participant is then asked their gender and male participants are asked if they are over 14 years old. These questions are necessary for the purpose of finding the best fitting vocal timbre, which is female for all participants except for male participants aged 14 and older.

Recordings are made via the Internet over a secure connection using the speech-recognition functions on web browsers. At present this capacity is only possible on Chrome, Firefox, and Opera; the user is alerted to this limitation on the SSAP home page. By consenting to participate in the SSAP, the user allows temporary microphone access while the SSAP runs; this access is stopped after the protocol ends, after the user closes the web page, or after an extended period of inactivity. For each recorded trial on the SSAP, the participant is asked to record again if there is no audio data of sufficient quality to extract a sung pitch.

(1) Warmup subtest

As the title implies, the purpose of this subtest is to warm-up the vocal folds and muscles that control pitch and respiration. Furthermore, responses to these tasks – which require no matching to a target pitch or imitation of a melodic pattern – provide an estimate of the comfortable singing range for a participant.

The participant first views a menu of familiar songs: *The ABC Song; Twinkle, Twinkle, Little Star; Mary Had A Little Lamb; Are You Sleeping (Frère Jacques)?;* and *Jingle Bells.* The participant selects a song by clicking on its title, and then sees instructions to sing it in a self-selected key while viewing lyrics. After clicking on a *Play* icon (indicating continuation), the lyrics appear below a flashing text-box that says *Record*, indicating that SSAP is now recording the participant's vocalizations. A *Stop* icon appears below the lyrics, which the participant clicks when he or she wants to stop the recording. The songs on the response menu were chosen based on their familiarity and simplicity. After some consideration we decided not to include *Happy Birthday to You* due to its complexity.

In the second warmup task, the participant is asked to produce a *comfort pitch*, defined simply as a single pitch the participant finds comfortable to sing. The participant is asked to sustain this single pitch on the syllable "doo" for a couple seconds. As with the familiar song task, the participant clicks an arrow symbol, which is followed by a screen showing the flashing *Record* box above the stop icon. Although the notion of a comfort pitch is highly subjective, many researchers use this kind of task to estimate comfortable singing range and it appears to be the best possible way to estimate the comfortable singing range for an individual.

After both warmup tasks are recorded, and saved as digital .wav files, SSAP concatenates both files, extracts the fundamental frequency (f_0) of the voice throughout, and forms a histogram of f_0 values. The median f_0 from this distribution is then used to identify a comfortable key to use in later vocal imitation trials. Specifically, SSAP selects one of five possible keys for which the mediant (i.e., major 3rd) scale degree is close to the participant's median f_0. The scale's mediant constitutes the midpoint of the pitch range used in imitation trials, where all pitches come from the first five scale degrees of a major scale. In addition, for imitation trials we match the timbre of vocal models to participants based on gender, and this further constrains key choice. Male vocal timbres can map to major keys based on A_2, C_3, D_3, F_3, and A_3; whereas female vocal timbres can map to the keys based on F_3, A_3, C_4, D_4, and F_4. Analyses of our database so far, reported later, suggest that these key choices are appropriate for our participants.

(2) Pitch matching to a vocal model

After the comfortable range is detected, participants are given the first of three subtests that involve imitative singing in a call-and-response format. These three subtests constitute the core of SSAP both because imitative singing is easier to score reliably, given the presence of an explicit target, and because these tasks are the most widely used in the existing literature on singing accuracy. The first subtest begins with a practice trial using a pitch that does not appear on any other trial (a minor 3rd above the tonic of the key selected for the participant). Participants can repeat this practice trial as many times as they wish.

Each trial in this subtest is initiated by the participant as was done for the warmup trials (by clicking the *Play* icon). Immediately thereafter the participant hears the target stimulus, followed immediately by a flashing text-box that says *Record* positioned above the *Stop* icon. The participant is instructed to listen without singing along to the target, and then to sing back as soon as possible when the visual *Record* cue starts. After the participant stops recording, SSAP's analysis procedure (described later) begins automatically and runs in the background while future trials are run.

Vocal targets were produced by two university voice students, one male and one female, instructed to sing with minimal vibrato. Each student produced several instances of each sung note on the syllable "doo" for a wide range of pitches. Targets were selected from these initial recordings based on the stability and accuracy of sung pitches, and were occasionally edited further for intonation accuracy based on equal tempered tuning.

Participants complete 10 trials of vocal pitch matching, comprising the first 5 scale degrees of their chosen key. Each pitch is repeated twice, with all five pitches appearing in a random order for the first five trials and then in a different random order for the next five trials.

(3) Pitch matching to piano tones

In many respects this subtest is identical to the previous one. The primary exception is that target stimuli are piano timbres. The pitches used in this subtest are identical to those used in the previous subtest, but the 10 trials are ordered differently.

The rationale for this subtest comes from several previous studies suggesting that singers are sensitive to timbre, and imitate vocal timbres more accurately than instrumental timbres, with this difference being enhanced for less accurate singers (Hutchins, Larrouy-Maestri, & Peretz, 2014; Hutchins & Peretz, 2012; Lévêque, Giovanni, & Schön, 2012; Moore, Estis, Gordon-Hickey, & Watts, 2008). Vocal imitation of piano timbres in this subtest may thus be contrasted with performance on the previous subtest.

(4) Imitative singing of 4-note vocal patterns

The final imitation subtest involves having participants listen to a 4-note pattern based on the same vocal timbre used in the first imitation subtest and then sing it immediately afterwards. Following a single practice trial that can be repeated as often as needed, 6 experimental trials commence that use patterns that differ from the practice trial. Each pattern comprises 4 different pitches selected from the first 5 scale degrees from the participant's key. Melodies begin either on the tonic (3 trials) or the dominant (3) trials, and vary in melodic contour. Melodic patterns were created by splicing together single pitches sung by vocal models. Each tone is approximately 900 milliseconds and is followed by a pause so that inter-onset intervals are 1 second long. Every pitch in the pattern is scored separately for a possible score of 24.

(5) Familiar song singing

Next, participants are asked to sing the same familiar song that they sang during the warmup subtest (1). First, they are asked to sing using the song's lyrics, as they did before, but, unlike the warmup, this recording is used for measurement of singing accuracy. Then participants are asked to sing the song again using the syllable "doo" for each syllable of text, in keeping with research suggesting that the accuracy of song-singing is influenced by whether participants have to recall text information while singing (Berkowska & Dalla Bella, 2009; Racette & Peretz, 2007).

(6) Simple pitch discrimination

Several studies to date have investigated associations between singing accuracy and pitch discrimination, with varying results (e.g., Dalla Bella, Giguère, & Peretz, 2007; Demorest, 2001; Moore et al., 2008; Pfordresher & Brown, 2007; Wise & Sloboda, 2008). For the SSAP we chose a psychophysical staircase approach to pitch discrimination similar to one of the leading studies concerning this issue (Loui, Guenther, Mathys, & Schlaug, 2008). On each trial, the participant hears two sine tones in succession and reports whether the second tone is higher or lower than the first by clicking one of two vertically oriented response boxes. The first tone in the sequence is always the same, and the second tone can be higher or lower with equal probability, with the magnitude of the difference varying adaptively. Specifically, the difference between the tones grows smaller by a factor of ½ after three successive correct trials, or larger by a factor of 2 after a single incorrect trial (this is referred to as a the 3-up, 1-down adaptive staircase procedure). After 6 reversals in the direction of change (harder to easier, or the reverse), the task ends and the final pitch difference is reported in both Hertz and cents as the participants' threshold.

(7) Questionnaire responses

Finally, the participant is asked to provide information concerning his or her musical background, including when they took their last music class, whether they received any formal training on instruments or voice, and whether they participated in a music group of any kind. For all yes responses, participants are asked to indicate the number of years of study or music participation. Participants are also asked to self-evaluate their interest and ability as singers on a 7-point Likert scale from *Strongly Agree* to *Strongly Disagree* by responding to the following statements. 1) I enjoy singing, 2) People think I am a good singer, 3) I am musically talented. Finally, we ask about participants' linguistic experience and the country in which they are taking the SSAP.

The SSAP analyzes accuracy of participants' singing and pitch discrimination in the background while the tasks progresses. Participants receive feedback on their performance immediately after finishing the protocol. Accuracy in the imitative tasks is reported as % of pitches sung within +/- 50 cents of the target pitch, a metric that is easily understood by the general public. Accuracy of song singing involves a percent of match between the frequency histogram of produced pitches to an ideal histogram and pitch discrimination thresholds are displayed using both Hertz and cents measures. We describe the details of these scoring procedures in the next section.

On-line acoustic measures of singing

Pitch extraction and range-finding

The SSAP is presently implemented as a series of Matlab routines (the Mathworks, Natick, MA), which are initiated by Perl scripts that form an interface between Matlab and the Internet. As such, we opted to extract pitch using the Matlab-based algorithm Yin (de Cheveigné & Kawahara, 2002). This algorithm uses the autocorrelation method of f_0 extraction with additional constraints based on musical applications (including but not limited to singing) that are designed to prevent underestimation of f_0 for contexts in which high pitches frequently occur. The output of Yin includes estimates of f_0 for each sample as well as the amount of harmonicity and spectral power, which can be used to separate accurate from spurious f_0 estimates. During the initial development of the SSAP we determined settings that yielded reliable f_0 estimates in a range of recording environments and for voices of varying pitch heights.

Automated scoring of accuracy for single pitches

All pitch imitation tasks involve estimating the difference between target pitches and imitated pitches on a note-by-note basis. For the single-pitch imitation tasks this involves comparing the entire sung reproduction to the entire sung target. The challenge in each case involves extracting a point estimate that reflects sung pitch independent of any “scoops” of pitch at the beginning and/or end of notes as well as outlying pitch estimates that may results from problems of pitch extraction. We have found that the best resolution to this problem is to limit pitch analyses to the central portion of pitches (the inter-quartile range of all extracted samples), and to use median f_0 rather than the mean, in order to limit the influence of outliers. This same procedure is applied to both the target recording and the recorded imitation by the participant.

The SSAP records the signed and absolute differences between target and imitated pitches in cents (where 100 cents = 1 semitone) in an output file that is available to the researcher. Output to the user is based on categorizing these differences as reflecting accurate or inaccurate pitch matching, where any absolute difference greater than 50 cents is considered an error. User output is based on the percent of sung pitches that are categorized as accurate or inaccurate, out of the 10 trials for each subtest involving single pitch imitation.

When extracting f_0 from an audio recording, it is common for algorithms to estimate pitches from the wrong octave. Often these octave artifacts can be abrupt and transient, occurring within a sung note. The aforementioned use of the median rather than the mean helps avoid the potentially biasing effect of these occasional outliers. Nevertheless, we also incorporated an octave correction procedure within each sung pitch, based on advice from Hutchins (personal communication). The algorithm is based on a ‘smoothness tolerance’, which is the kind of pitch transition that is considered allowable between two adjacent samples (about .03 milliseconds given our 32 kHz sampling rate). We set this tolerance to be a minor third (300 cents), following Hutchins’ lead. For every pair of samples exceeding this threshold, the second sample was adjusted either up or down an octave depending on the direction of the transition.

Note segmentation for pitch pattern imitation

Although human listeners find it relatively simple to perceive where one note ends and the next begins, these boundaries are not obvious in the acoustic signal. Because we do not require singers to pause between sung tones, which could be unnatural and lead to insufficient pitch information, silences do not reliably indicate boundaries. Also, because many singers are inaccurate, algorithms based on dynamic time warping of an ideal melody to match the sung melody do not work well. Many singers in the SSAP database produce incorrect pitches, and fluctuations of f_0 within a sung note may be as large or even larger than f_0 transitions between notes for some singers. We here describe a procedure that has been used in the SSAP so far, and may be useful for other researchers as well (we can provide code on request). However, as discussed in supplementary material for this chapter, further analyses suggests that a radically simpler approach may be sufficient and computationally less expensive, thus providing an attractive alternative for future versions of the SSAP.

The algorithm for detecting note onsets to date uses fluctuations in the intensity of the acoustic signal that are associated with syllabification. This involved making a minor constraint on singing, which was that all notes should be sung on the syllable “doo”, given that the stop consonant /d/ provides a salient marker of the syllables beginning. Specifically, /d/ generates an abrupt increase in the amplitude envelope of the signal. When one takes the first derivative (velocity) of the amplitude envelope, this abrupt increase in intensity leads to a spike. The SSAP computes the amplitude envelope from voltage fluctuations in the wav file, that are then smoothed using a using a moving average window.[1] The algorithm then finds all of the peaks in velocity of dB, and then winnows these down to a subset of peaks that are above a certain threshold (based on preliminary analyses) and are associated with sampled f_0 information (spurious peaks can occur from changes in room noise). Finally, a dynamic search routine is run that identifies the first candidate onset, then implements a refractory period before finding the next candidate onset, and so on. The refractory period was instituted in order to prevent the identification of spurious note onsets in close proximity around a single syllable onset.

This onset detection algorithm can identify the timing of note onsets with high precision. Even if there are errors in the identification of note onset timing, our procedure for analyzing pitch, described above, allows a generous margin of error given that only the middle 50% of sample pitches are used. Nevertheless, the onset detection algorithm is not foolproof and sometimes fails due to the open-ended nature of online data recording. For instance, if recording occurs in a room with much ambient noise, extracted f_0 may be accurate but intensity fluctuations may be unreliable. Also, some participants do not articulate /d/ very clearly. As such, we adopted a second-pass solution based on hard-wired information about sung patterns. Specifically, if the algorithm fails to identify 4 onsets, it assumes that missing onsets are isochronously timed with approximately 1 second per onset, and imputes estimated positions of these missing onsets. Of course, the imputed onsets are not identified with as much accuracy as those resulting from the algorithm itself, but the procedure for identifying pitch, described earlier, makes this limitation unproblematic most of the time.

Taken together, the SSAP algorithm for identifying tone onsets is robust and in the vast majority of cases leads to accurate estimates of pitch accuracy. Nevertheless, those using the SSAP for research purposes should check the analyses and re-analyze trials if need be. The SSAP output includes plots of every trial showing where notes are segmented that researchers can consult in order to evaluate the accuracy of note segmentation and pitch analysis.

Automated scoring of familiar song singing

A deeper challenge arises when one attempts to analyze the accuracy of song singing. Song lyrics do not include reliable acoustic cues for the beginnings of syllables, which may start on vowels, liquids or other sounds that do not involve a burst of intensity. Even when singing on the syllable “doo”, familiar songs often include rhythmic variability that makes singing difficult to process, relative to imitative singing.

Based on these challenges, our solution in the SSAP is to analyze song accuracy based on statistical properties of f_0 throughout the sung sequence. Specifically, we generate a histogram of f_0 values, after quantizing each sampled f_0 to match the nearest possible semitone. Rhythm plays some role in this frequency distribution in that the frequency with which each pitch class appears is weighted by the summed duration associated with that pitch across the song, similar to the practice of using duration-weighting in generating tonal profiles (Smith & Schmuckler, 2004). However, this method sidesteps the problem of trying to identify specific note boundaries, mentioned before.

The f_0 histogram from the imitation is then correlated with a histogram representing the distribution of f_0 under note-perfect conditions. The pitch classes in this idealized histogram are adjusted in order to achieve the best fit to the participant's histogram, in order to accommodate different key choices by the participant. The r_2 from this comparison reflects the degree of match between an ideal performance and the participant's performance.

Our initial analyses of this algorithm indicated that it correlated with a measure of performance based on segmenting pitches into notes and judging the accuracy of produced intervals. However, the histogram-matching algorithm of the SSAP also systematically underestimated performance relative to the more traditional procedure. Because the SSAP is used by the general public, and may contribute to one's self-image with respect to singing, this underestimation was a serious concern. As such we introduced an adjustment that constrained match scores to vary between 40% and 100% correct according to an exponential function. This function was found to lead to a closer relationship between the histogram-matching algorithm and analysis based on note segmentation.

This method for scoring song singing was formulated in the knowledge that it would need revision after further data collection. Users of the SSAP are warned at the end of the protocol that the song-singing measures are under development. In the next section, where we report validity tests of the SSAP scoring procedures, we evaluate the validity of this initial method. In the supplementary material to this chapter we discuss a possible alternative approach for future versions of the SSAP.

Validity tests

In order to test the validity and robustness of SSAP's online scoring measures, we ran the automated procedures on several data sets that were scored using traditional offline acoustic techniques that are highly accurate yet more time-consuming and not practical for the online measure. The SSAP analyses are considered valid insofar as they match these other measures. In separate sections we consider the validity of online scoring for imitation tasks and for song singing.

Automated scoring of imitation tasks

The automated scoring procedures we run are designed to replicate the process of manual analyses done by hand. Often these analyses are done by segmenting a spectrogram by hand, using visual and auditory information to identify each individual onset. Analysis of the data sets reported here automated much of this process by using the onset detection algorithm described above, but critically these "by-hand" analyses involved having the experimenter check each onset afterwards by using auditory and visual information. It is unlikely that the automated procedures used in the SSAP could match such careful time-consuming practices in every case. In particular, we predicted that matches between the SSAP and scoring by hand would be most common under ideal recording conditions.

As such, we ran the SSAP's automated scoring algorithm on two data sets that reflect distinct recording environments. One data set was recorded in the first author's lab as part of a project that evaluated the role of laryngeal and facial muscles during singing and auditory imagery (Pruitt, Halpern, & Pfordresher, 2019). This study included 4-note novel melodies that were sung imitatively, much like the 4-note imitation task from the SSAP. Most important, recordings were made in a Whisperroom sound-attenuating booth using a high-quality dynamic microphone. Participants were college students (mostly non-musician) who were instructed by the experimenter concerning appropriate breath control, and were encouraged to sing loudly and articulately. Thus, we can expect that this data set should yield audio files that are most amenable to automated scoring.

The other data set was collected in a classroom setting via a microphone placed approximately 12 inches from the mouth of Kindergarten children (Demorest, Nichols, & Pfordresher, 2018). Recordings featured a great deal of ambient noise, including reverberation, occasional chair scrapes and occasional extraneous voices. Also, kindergarteners often sang much more quietly than the college students. Thus, this second dataset represents a non-ideal recording environment and a challenge to the automated scoring procedure.

Table 27.1 summarizes the comparisons between SSAP's automated scoring procedure, and analyses done by hand, which serve as the primary analyses for each data set. Figures in the "Lab" column refer to data from Pruit and colleagues (2019), whereas "Classroom" refers to Demorest and colleagues (2018). For both data sets, a large number of recordings were collected, with a recording constituting a complete 4-note pattern. First, we consider how often SSAP is able to generate some kind of accuracy score (completed analyses). This is reflected in the percent of completed SSAP analyses in Table 27.1. The primary reason for failed analyses is that the note segmentation fails to yield an equivalent number of notes to compare against the target, such as when one or more notes is omitted or the quality is poor enough that no pitch trace is recovered. Both situations are more prominent in the classroom dataset. As can be seen in Table 27.1, SSAP successfully generates an estimate of accuracy in virtually all trials under ideal lab contexts (99%), and the majority of time (although far less often, 88%) in the classroom. We hasten to add that most recordings from the general public are not as poor as in the classroom. Most individuals use the SSAP alone in a quiet room with minimal reverberation, due presumably to the presence of carpets, curtains and seat cushions.

Table 27.1: Comparison between SSAP scoring of imitative performance with "traditional" analyses computed by hand based on two recording environments.

	Lab	Classroom
Number of recordings	1,216	1,082
Completed SSAP analyses (%)	99%	88%
Number of analyzed notes	3,080	2,780
Pitch error rate, SSAP coding (%)	45%	59%
Pitch error rate, coding by hand (%)	45%	68%
SSAP / by hand coding match (%)	89%	86%
M pitch deviation, SSAP coding (cents)	99.47	285.69
M pitch deviation, by hand (cents)	92.52	144.66
M pitch deviation difference*	6.95	141.03
Difference within 100 cents (% notes)	97%	78%

* positive value means a higher deviation score for SSAP, in cents (100 cents = 1 semitone)

Although the number of trials SSAP can analyze at all is an important metric, a more important assessment of scoring validity involves comparing the match between the automated procedure and scoring by hand for those trials that are successfully analyzed. We focus on the two primary measures of singing accuracy: the percent of sung notes coded as error, and the mean deviation of sung from target pitches in cents.

For the purpose of these analyses the unit of analysis is a single note rather than a 4-note trial. As such in Table 1 we report the total number of notes that were analyzed, which is constrained in part by how many recordings were successfully analyzed.

As can be seen, there is a high degree of match between the pitches coded as an error by the SSAP metric and those coded as an error for analyses done by hand. The percent of notes yielding the same coding (error or accurate) was nearly 90% under ideal recording conditions, and were only slightly less prevalent under non-ideal conditions (Table 1). An important follow-up question concerns whether the percent of mismatching notes constitutes a bias within the SSAP measure. Ideally, mismatches would vary randomly between false positives (an accurate note coded as error) and false negatives, leading to a similar overall error rate arising from the SSAP analysis as well as analyses by hand. This is what was found for recordings under ideal conditions: The mean pitch error rate that resulted from SSAP's automated procedure was equivalent to the error rate in analyses done by hand (45% in both cases, a fairly high error rate). Under non-ideal conditions, however, a bias was apparent, with SSAP underestimating the prevalence of errors relative to computations performed by hand. This probably reflects the octave correction procedure in the SSAP, which may have detected apparent octave errors that were in fact indicative of poor-pitch singing.

With respect to mean absolute pitch deviations, the SSAP measure on average fell within 7 cents of analyses done by hand for recordings under ideal conditions. This is a difference that is considerably smaller than listeners' ability to detect mistunings in singing, which is probably at least 25 cents (Hutchins, Roquet, & Peretz, 2012;

Larrouy-Maestri, 2018; Pfordresher & Brown, 2017), and thus constitutes highly accurate performance. Likewise, the percent of pitch deviations from the SSAP that fall within 1 semitone (100 cents) of the corresponding pitch deviation found in analyses by hand was 97% for these recordings. Not surprisingly, recordings done under non-ideal conditions did not fare as well. The mean difference between SSAP and analyses by hand was over a semitone, with SSAP overestimating the magnitude of pitch deviations on average (similar to the over-estimation of error rates). However, this figure is influenced by outliers; the percent of pitch deviations across measures that fall within a semitone is still quite high (78%).

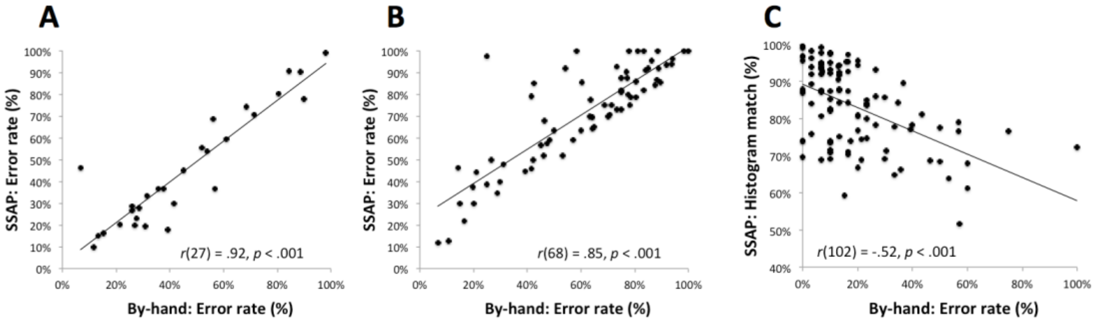

Figure 27.1: Scatterplots illustrating correlations between SSAP’s automated scoring and analyses conducted by hand, for vocal imitation tasks recorded under quiet conditions (A), for vocal imitation tasks recorded in a classroom (B), and for song-singing recorded under quiet conditions (C). See the text for further details.

We also examined how well individual differences in the SSAP scores correlated with individual differences in scores based on traditional analyses. For these correlations, we first aggregated across all trials for each participant (N = 29 for recordings in an ideal environment, N = 70 for recordings in the non-ideal environment). Figure 27.1 shows scatterplots for correlations across error rates derived from SSAP’s automated analyses and error rates based on analyses by hand (panels A and B). All correlations were one-tailed, given that negative associations would indicate failure of the SSAP scoring procedure. For recordings under ideal conditions (Figure 27.1A), the correlation across error rates was strong and positive, $r(27) = .92$, $p < .001$. Correlations for recordings carried out under non-ideal conditions (Figure 27.1 B)were lower, but still strong and significant, $r(68) = .85$, $p < .001$. In general, these analyses reveal that automated analysis of singing accuracy based on more coarse-grained categorical measures, such as error rates, may be more robust than analyses that require greater accuracy, such as pitch deviation scores.

To summarize, although the validity of the SSAP's scoring of vocal pitch imitation is influenced by recording environment, it was found to be surprisingly robust. The non-ideal recording environment here probably constitutes the worst possible environment in which one could reasonably analyze singing (assuming people do not try to run the SSAP in a crowded train station). Moreover, most users from the general public would likely run the SSAP in a quiet private room to avoid possible embarrassment. Thus, based on these data we consider the automated scoring of vocal pitch imitation in the SSAP to be valid particularly for recordings made for research purposes in a quiet environment.

Automated scoring of familiar songs

Next we report validity tests for automated scoring of familiar songs. As noted earlier, these performances present challenges for automated analysis that are either absent or far reduced in imitative performances. As such, SSAP scoring to date is based on the approach described earlier, which we refer to as the "histogram match" approach. We here evaluate the success of this initial approach; a potential alternative approach for future versions of the SSAP is discussed in supplementary material.

We report an analysis of 103 recordings of either *The ABC song* or *Twinkle, twinkle, little star* (the same melody). There were 61 participants who sang the song once using lyrics and once on the syllable "doo"; five recordings were omitted because the performance was not complete. The sample was collected by Sean Hutchins for an independent project and shared with us for the purpose of this validation exercise. Recordings were carried out under quiet conditions, and performers included non-musicians, instrumentalists, and vocalists.

We tested the validity of SSAP's automated scoring of familiar songs by correlating this measure with a measure of accuracy based on segmenting individual notes by hand and assessing the accuracy of sung pitch intervals. The resulting correlation was significant and negative, $r(102) = -.52$, $p < .001$, which was anticipated given that the SSAP score reflects degree of match to the frequency distribution of pitch in the model (see previous section). Although this result does offer validation for the SSAP procedure, the effect size is far less than what we found for validity tests of the automated procedure for scoring imitative singing, accounting for a fairly small portion of variance ($r_2 = .27$). Thus, in keeping with the proviso included on the feedback screen of the SSAP interface, these scores should be treated with caution. Future versions of the SSAP may incorporate the revised routine described in the supplementary material. We also plan to present participants with percentile ranks based on our existing database (over 1,000 data sets), to which will help users better interpret the meaning of their scores. We are presently analyzing performance from the existing database for future publication.

Conclusion

We have provided a detailed technical introduction to the SSAP that complements the earlier conceptual introduction (Demorest et al., 2015), including technical details of the interface and scoring procedure. We have also presented evidence for validity of the automated scoring measures. At present, scoring of imitative singing for short items (1-4 notes) is highly accurate whereas scoring of song singing has considerable room for improvement. As such, we will continue to recommend that users focus primarily on scores for imitative singing. Given that SSAP provides researchers with .wav recordings of every test item for further analysis.

The SSAP is available to researchers in addition to the general public. The tool can be used as a screening test or a measure of singing accuracy as part of a larger experiment. We recommend that researchers take advantage of the detailed analyses that the SSAP offers as opposed to including only the feedback given to participants at the end of the SSAP. Researchers may also want to check the automated analyses using the recorded audio files that are produced by the SSAP.

We think this protocol has great promise as a standardized measure of singing accuracy. It is short enough (total running time approximately 20 minutes) to allow researchers to use it in combination with any specialized measures they wish. Ideally, the SSAP can facilitate standardization across research studies, allowing cross-comparisons and meta-analyses. In addition, we hope the general public may use the SSAP to get a more objective sense for their own singing accuracy and use it as a guide to facilitate and better enjoy their skill in singing.

Acknowledgement

The development of the SSAP was funded by a grant from the National Association for Music Education. We thank Göktuürk Poyrazoğlu for assistance in programming, and Sean Hutchins for providing recordings of piano tones and analyses of sung performances. Second author Steve Demorest tragically passed away in the interim between initial submission and revision. The first author therefore undertook all revisions, with the hope that the final product may do justice to the Steven Demorest's signficiant contribution to this area of research.

References

Berkowska, M., & Dalla Bella, S. (2009). Reducing linguistic information enhances singing proficiency in occasional singers. *Annals of the New York Academy of Sciences, 1169*, 108-111.

Berkowska, M., & Dalla Bella, S. (2013). Uncovering phenotypes of poor-pitch singing: The Sung Performance Battery (SPB). *Frontiers in Psychology, 4,* 714.

Cohen, A. J. (2015). The AIRS Test Battery of Singing Skills: Rationale, item types and lifespan scope. *Musicae Scientiae, 19*, 238-264.

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121*, 1182-1189.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*, 1917-1931.

Demorest, S. M. (2001). Pitch-matching performance of junior high boys: A comparison of perception and production. *Bulletin of the Council for Research in Music Education, 151*, 63-70.

Demorest, S. M., Nichols, B. E., & Pfordresher, P. Q. (2018). The effect of focused instruction on young children's singing accuracy *Psychology of Music, 46,* 488-499.

Demorest, S. M., & Pfordresher, P. Q. (2015). Seattle Singing Accuracy Protocol - SSAP [Measurement instrument]. https://ssap.music.northwestern.edu/.

Demorest, S. M., Pfordresher, P. Q., Dalla Bella, S., Hutchins, S., Loui, P., Rutkowski, J., et al. (2015). Methodological perspectives on singing accuracy: An introduction to the special issue on singing accuracy (Part 2). *Music Perception, 32*, 266-271.

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception & Psychophysics, 76,* 2522-2530.

Hutchins, S., & Peretz, I. (2012). A frog in your throat or in your ear? Searching for the causes of poor singing. *Journal of Experimental Psychology: General, 141*, 76-97.

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception, 30*, 147-159.

Larrouy-Maestri, P. (in press). “I know it when I hear it”: On listeners’ perception of mistuning. *Music & Science.*

Lévêque, Y., Giovanni, A., & Schön, D. (2012). Pitch-matching in poor singers: Human model advantage. *Journal of Voice, 26*, 293-298.

Loui, P., Guenther, F. H., Mathys, C., & Schlaug, G. (2008). Action-perception mismatch in tone-deafness. *Current Biology*, *18*, R331-332.

Moore, R., Estis, J., Gordon-Hickey, S., & Watts, C. (2008). Pitch discrimination and pitch matching abilities with vocal and nonvocal stimuli. *Journal of Voice, 22*, 399-407.

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception, 25*, 95-115.

Pfordresher, P. Q., & Brown, S. (2017). Vocal mistuning reveals the origins of musical scales. *Journal of Cognitive Psychology*, *29*, 35-52.

Pruitt, T. A., Halpern, A. R., & Pfordresher, P. Q. (2019). Covert singing in anticipatory auditory imagery. *Psychophysiology, 56*, e13297.

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition*, *35*, 242-253.

Smith, N. A., & Schmuckler, M. A. (2004). The perception of tonal structure through the differentiation and organization of pitches. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 268-286.

Wise, K., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined 'tone deafness': Perception, singing performance, and self-assessment. *Musicae Scientiae, 12*, 3-23.

Supplementary Material

We here discuss tests of alternate approaches to the SSAP's automated analyses procedures. These approaches are not presently part of the SSAP's architecture, but will likely be incorporated in the second version of the SSAP, which is presently under development.

Alternative note segmentation approach for imitation of short melodies

As discussed in the chapter, SSAP presently segments imitatively sung melodies into individual notes based on changes to intensity at the start of the syllable "doo". Although this procedure works well, it runs into occasional problems (e.g., if there is not a pronounced change to intensity at syllable onset), and involves considerable processing. Thus, we have tested a much simpler alternative approach based on simply dividing a sung melody into 4 equal sections. This decision was based on careful inspection of many plots of trials. Participants are usually successful at timing note onsets isochronously, and do not vary much from the absolute timing of target files. This, combined with the generous room for error afforded by our analysis of pitch accuracy (described in the chapter), made this simpler option feasible.

The results from this new procedure led to results that are identical to those described in Table 27.1 with respect to estimated error rates and pitch deviations, albeit with all notes analyzed. Furthermore, run time estimates from Matlab suggest that this new procedure may shorten the time it takes to analyze the accuracy of a pattern imitation in half. Based on this success, we anticipate using this simplified strategy in future versions of the SSAP. Users will likely notice faster processing time with similar validity in automated scoring.

Alternative automated scoring for familiar song singing

As noted in the chapter, validity tests of SSAP's automated scoring for familiar song singing indicated substantial room for improvement. This is not surprising given the considerable complexity of analyzing songs.

We tested a novel procedure similar to our testing of the simplified process of segmenting notes that we described earlier for imitative singing. We started by segmenting the sample melodies into 24 isochronous units. This involves treating immediately repeated pitches (e.g., "A, B", in *The ABC song*) as a single unit. This simplification allows us to minimize the number of potential mismatches between sung and target onsets. In a first-pass analysis we analyzed the success of this approach, but found several problems that related to the presence of pauses. In many cases, pauses would disrupt a produced rhythm because the duration of a pause would lapse beyond the expected onset for the subsequent segment in a melody. To correct for this we added a subroutine that identified all pauses (defined as gaps in f_0 longer than 500 samples, approximately 18 milliseconds), whose total duration spanned across the subsequent onset (i.e., the pause began after one segment and ended after the anticipated onset of the next segment). When this happened, all subsequent onsets were delayed in exact proportion to the lapse caused by the ending of the pause.

This additional subroutine improved the performance of our revised scoring procedure. The correlation of error rates estimated form this new procedure and error rates based on segmentation by hand was *r* = .64, reflecting an increase in variance accounted for from .27 to .41. This is a fairly large sized improvement; however, the change in *r* was not statistically significant according to tests for dependent *r* values (Cohen & Cohen, 1983).

However, this analysis misses an important point. Inspection of individual cases revealed an important shortcoming of the previous histogram-match approach: Monotone performances can actually yield a high match score. We discovered this after reviewing recordings from a middle school sample. One participant simply spoke the words to *The ABC song*, and yet achieved a match score from the current SSAP scoring of 73%! This happens because the algorithm matched the (approximately) single pitch the participant produced to the middle of the range of *The ABC song*, which ends up being a prevalent pitch class in the ideal histogram. Clearly this match score was a misrepresentation of the performance, even given the generous correction we added. By contrast, the new procedure led to a percent correct score of 9.5% (2 out of 21 notes produced correctly).

References

Cohen, J., & Cohen, P. (1984). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd Ed). Hillsdale, NJ: Lawrence Earlbaum Associates.

[1] Matlab code retrieve from http://music.columbia.edu/cmc/musicandcomputers/popups/chapter3/xbit_3_1.php % on 8 October, 2014. The link is presently inactive to date.