When Audiences Become Performers and Speech Becomes Music: New Tools to Analyze Speech, Song, and Participation in Chinese Crosstalk

Francesca R. Sborgi Lawson, Joshua David Sims and John Scott Lawson

Abstract
As a presentational performance in which two actors perform a scripted dialogue that includes several short vocal selections, Chinese crosstalk features animated exchanges between the actors while also inviting enthusiastic involvement from the audience in the form of cheering applause. Inspired by research on the musicality of spoken bouts between same-sex friends (Hawkins, 2014; Robledo Del Canto et al., 2016; Cross et al., 2016) the present study focuses on two selections in a crosstalk performance—looking at not only the vocal interactions between the actors, but also those between the actors and the audience. Using ELAN, a tool for analysis of video and audio, Praat acoustic analysis software, and R statistical software, the study reveals (1) long durations of audience responses, signifying the seminal importance of the audience as a third performing agent in Chinese crosstalk, (2) significant portions of the performance that demonstrate rhythmic entrainment between performers and audience members, and (3) highly significant examples of approximately matching between the ending sounds of utterances and the initial sounds of subsequent responses in the final selection of the performance.

Keywords
Audience response, Chinese crosstalk, ELAN, musicalized speech, Praat, presentational performance, R

Introduction
Chinese xiangsheng—often translated as crosstalk—is a comedic dialogue not unlike the repartee between Abbott and Costello in their classic routine, “Who’s on First?” As part of the shuochang or “speaking-singing” tradition in north China, whose very name confirms an interactivity between speech and song, crosstalk also contains short vocal performances in addition to comedic banter. Crosstalk dialogues are always scripted, follow a set format, and yet appear unhearsed. Pieces in the crosstalk tradition are conceived in a four-part form, climaxing when performers reach the baofu or “package,” which is the punchline of the piece (Lawson, 2011, p. 117). Tsau explains that the baofu amounts to catching an audience by surprise with a reasonable but unexpected outcome (1979-80, pp. 61–62).

The emphasis of the present study is the interplay between actors and audience and between speech and song, showcased in a video-recorded performance of “A Carefree Life,” featuring Guo Degang and Yu Qian, two of China’s most popular crosstalk actors. Published on February 15, 2016 on YouTube, “A Carefree Life” is a short, 14-minute clip that epitomizes Guo’s brash and irreverent style—with Yu Qian, his longtime collaborator and fellow actor, as his “straight man.”

Throughout Guo and Yu’s performance, Guo’s credibility is constantly questioned as he says ridiculous, nonsensical things. Consequently, when Guo eventually...
challenges Yu to start singing a notoriously high-pitched Beijing opera aria—whose even higher-pitched ending Guo declares he will sing as the finale—Yu is understandably nervous. Given Guo’s preposterior and irrational claims throughout the performance, will he be able to do it? In the final section leading up to the baofu, Guo mercilessly teases Yu, who begs Guo to start in a low key. By giving Yu the starting pitch in different registers—first much too low and then too high—Guo orchestrates the suspense in preparation for the climactic ending. The interplay between Guo and Yu as they battle over the key of the aria becomes the comedic energy that fuels the audience’s explosive responses during the final part of the performance.

Before examining the musical and participatory nature of this crosstalk, it is important to underscore the significance of the groundbreaking research on conversations between pairs of same-sex friends conducted at the University of Cambridge (Hawkins et al., 2013; Hawkins, 2014; Robledo del Canto et al., 2016) as the motivation behind the present study. Using ELAN, a tool for analysis of video and audio developed by the Max Planck Institute, the researchers at the University of Cambridge were able to look at pulse and intonation in conversations between eight pairs of same-sex friends, focusing on the turn transitions in which the second speaker picks up after the first speaker stops. The Cambridge group discovered that speech can become temporally aligned in spontaneous interaction when the two speakers were attitudinally aligned (Ogden & Hawkins, 2015). Cross et al. (2016) noted that rather than conceiving of conversation as “transmit-receive,” the recorded speech interactions seemed better understood as arising from processes of cooperative co-construction between speakers. Attitudinally-aligned question-answer pairs were organized around a shared pulse across turn transitions—a pulse that was derived from speech accent actualized as fundamental frequency ($f_0$) peaks in the speech stream. These question-answer pairs could also exhibit pitch alignment between the modal $f_0$ of the first speaker’s utterance and that of the second speaker’s response. One of the conclusions from the Cambridge study is recognizing that speech and song may both be underpinned by common processes in certain contexts that demonstrate attitudinal alignment between the speakers.

In addition to emphasizing the importance of studying speech and song comparatively, the Cambridge group also proposes that researchers focus on interactive, participatory speaking and music making, rather than presentational performances, when trying to study the biological underpinnings of music and language (Cross, 2014). In making the distinction between participatory and presentational performances, they build on work done by Thomas Turino (2008), who, in turn, draws on ideas proposed by Charles Keil (1987). Turino explains that participatory performance occurs when there are no distinctions between artists and audience members, involving a maximum number of people in some performance role. By contrast, presentational performance refers to situations where one person or group of people provides music for the audience, who do not participate in making the music or dancing to it (2008, p. 26). Turino acknowledges that sitting silently while listening to music in a concert hall is a kind of musical participation, but not one in which dancing, singing, clapping, and instrument playing are integral to the performance (p. 28).

Crosstalk is useful for understanding the value of participatory interaction because it is both presentational and participatory, although not exactly in the way Turino describes the two modes of interaction. In crosstalk, audience members, who remain seated at all times, buy tickets to witness the way the two actors perform their dialogue on stage—a dialogue written by an author who specializes in the genre. At the same time, the performance is also punctuated by spontaneous applause, laughter, and other affiliative vocal sounds from members of the audience. In other words, crosstalk invites the audience, as a collective body, to participate by making vocal utterances throughout the presentational performance, functioning essentially as an interlocutor. As Clayton (2007), Clayton and Leante (2015), and Leante (2017) underscore the centrality of audience participation in the performance of North Indian Khayal, so audience participation is similarly indispensable to the success of a crosstalk performance. One could easily argue that audience participation effectively makes crosstalk a triologue rather than simply a presentational dialogue featuring the two main actors.

However, in comparing the Cambridge study with the crosstalk project, there are some major differences in performance and methodology. First, crosstalk performance is scripted, and therefore unlike the spontaneous interactions between pairs of same-sex speakers in the Cambridge study (Cross et al., 2016; Hawkins, 2014; Robledo del Canto et al., 2016). Nevertheless, presentational crosstalk performances elicit loud and exuberant vocal responses from the audience throughout the performance—responses that are spontaneous and participatory. Consequently, we suggest that participatory and presentational elements are not mutually exclusive, but rather complementary and scalable aspects of performance.

Second, the spontaneous utterances of the audience—vis-à-vis the scripted utterances of two performers—were substantial. In the sections about duration data from Bouts 1 and 2 of the crosstalk, we show that the durations of spontaneous audience utterances—as cheering applause and laughter—were as long as, and sometimes longer than, the utterances made by the performers, demonstrating that audience utterances constitute a significant part of the performance.

Third, when determining pulse in the present study, we had to quantify periodicity differently than the way it was done for English speakers in the Cambridge study (Robledo et al., 2016). Since Chinese is a syllable-timed language, as opposed to English, a stress-timed language
(J. Y. Chen et al., 2003; Chen & Chen, 2012; Roach, 1982), we calculated the pulse between all syllables in a given utterance—a different method than the one employed by the Cambridge study in which measurements were based on the time interval between stressed syllables that were marked by pikes in pitch. The analysis of stress is in Chinese is a difficult subject, evidenced by the conflicting opinions regarding the way stress interacts with tone and other factors (Chen, 1993; Chrabaszcz et al., 2014; Shen, 1993; Tseng & Lee, 2004). Nevertheless, regardless of where stress may fall in Chinese, any pattern in periodicity will appear between all syllables, necessitating a categorically different methodology for quantifying periodicity in the xiangsheng (XS) study. Despite the disparities between the two languages and the different procedures employed in the two studies, the XS study determined a certain degree of periodicity between the two performers and between the audience vis-à-vis the performers. A more detailed explanation about the way periodicity functioned in the XS performance follows in the section on pulse under Bout 2.

Fourth, periodicity functions differently in a crosstalk performance than in the spontaneous interactions between speakers in the Cambridge study because of the nature of a presentational crosstalk performance. In the same way that audience gestures and vocal responses in 

khvål

performance are constrained by the performance event (Clayton, 2007, p. 91), so audience response is constrained by the actors’ performance in crosstalk. Consequently, the presentational nature of the performance provides a reason why the audience reaction to crosstalk does not demonstrate regular periodicity throughout the entire performance, even though significant portions of the analyzed sections demonstrate periodicity among all three performing agents (see section on pulse under Bout 2).

Moreover, another reason for the absence of a regular pulse during parts of the crosstalk performance is suggested by Hawkins (2014), who claims that “when speech is used to convey referential meaning, then the words, and their sequential order, are largely dictated by the language… [and] strict rhythmic cycles may have to take second place to intelligibility” (5). Since the meaning and execution of presentational speech is the focus of the performance in crosstalk, the intelligibility and performative appeal of the actors’ speech sometimes overrides the periodicity that might otherwise occur in everyday spoken interactions.

Finally, the crosstalk performance demonstrates a tendency towards pitch alignment between the final pitch of an utterance and the first pitch of the response for all three agents—the two crosstalk actors and the audience—in the final bout of the performance (Bout 2), suggesting a phenomenon similar to the pitch contour matching found in Gorisch et al.’s study (2012). The authors in that study discovered that aligning insertions were significantly more similar to the immediately preceding turn in terms of pitch contour than were non-aligning turns, demonstrating affiliation with the previous speaker’s agenda (2012, p. 74). This finding suggests a similar occurrence to the pitch-approximating phenomenon discovered in the turns between the actors and between the actors and audience late in the performance of the crosstalk—an emotionally-charged time when performers and audience members demonstrate strong affiliation with the previous agent’s agenda. Even though Chinese is a tonal language, Levow (2019) has demonstrated that intonational cues quite similar to those in English are also employed in Chinese with lower pitch and intensity at the ends of turn units than at the beginnings of those turn units, signifying the importance of intonation over lexical tone in turns (77). Additionally, Nissen (2019) explains that the distribution of lexical tones occurs randomly over this sample of utterances, creating a mixture of tones. Although lexical tone is a slight driver of fundamental pitch, the overall pitch register is what matches among all three agents (the two performers and the audience). Finally, the audience response is not driven at all by lexical tone, and yet the pitch-aligning phenomenon still occurs, suggesting that pitch approximation and entrainment across turns happen regardless of lexical tone.

The following research and analysis employ ELAN, Praat acoustic analysis software, and R statistical software. Using these tools, we analyze the symbiotic relationship between presentational and participatory aspects of performance and study the reciprocal and synergistic connections between music and speech by demonstrating the different ways audience members interact vocally with performers in Chinese crosstalk. We focus on three major points in this paper: (1) the ways in which the audience becomes a third performing agent in a presentational setting, (2) the tendency towards rhythmic entrainment between performers and audience members for approximately 50% of the analyzed sections of the performance, and (3) the tendency for pitch approximation in the final bout of the crosstalk performance as a result of increased affiliative involvement between audience members and performers—an example of what we refer to as “attitudinal alignment.” For all three points, we argue that performer-audience interactions and speech-music relationships are best seen as part of a continuum rather than as opposing and dichotomous elements.

Two Selections or “Bouts” from the Crosstalk Performance, “A Carefree Life”

Given the Cambridge researchers’ experience in analyzing conversations between pairs of same-sex speakers, we decided to use ELAN to focus on pitch and duration in two bouts in the crosstalk recording, analyzing the vocal interactions between Guo and Yu and between the audience and the two actors. However, it should be noted that there was one important difference in the ways the two studies selected the bouts. In the Cambridge Study the researchers focused on question-answer pairs, finding that attitudinal alignment could be marked by the production and
pragmatic use of pulse across turn transitions. Instead, the XS researchers selected two bouts—one at the beginning and one at the end of the performance—to see if there were differences in periodicity and pitch during these two very different points in the performance.

The first bout in the XS study is entirely spoken, with no singing (from minute 1:00 to minute 2:14), featuring Guo’s musings on his future as a 140-year-old performer with Yu. The improbability of living, not to mention performing, at such an advanced age is an example of the kind of wild exaggeration that is expected in a crosstalk performance (Li, 1985). The second bout showcases both men taking turns singing a difficult opera aria (from minute 12:50 to minute 14:57) at the end of the performance, highlighting an exchange about singing in falsetto—representing the climax of the piece. Fascination with male actors singing in falsetto is part of the tradition of Beijing opera, and its particular difficulties and challenges for male singers can easily become a subject of levity in crosstalk, a genre known for its no-holds-barred practice of extravagant and hilarious satirical play.

**Tools for Analysis: ELAN and Praat**

ELAN—developed and maintained by The Language Archive (TLA) since 2000 and intended for use in language documentation and description—is a tool for creating complex annotations on video and audio resources. ELAN’s analysis suite can also be effectively used to annotate recordings of a quasi-musical nature, such as the recordings used in the Cambridge study; mixed spoken and sung performances, such as the crosstalk recording analyzed in this paper; or musical recordings.

The primary features of ELAN relevant to this project are its support for multilayer, hierarchical annotations, as well as the synchronization and integration of video media. The annotation system enabled the construction of annotations first at the phrasal level, then at the word level, as well as allowed for the attachment of additional tiers of annotations containing transliterations (Figure 1). The phrasal annotations allowed for the consideration of relative speaking time per participant and for the study of pitch approximations during bouts. The word-level annotations were used to generate data such as a map of the overall emergent melody of the spoken portions. The close synchronization of media playback with the transcription timeline in ELAN was critical for accurate annotation, as consultation with the video allowed for disambiguation when the identity of the speaker of an utterance was unclear in the audio recording.

For the selected crosstalk performance, ELAN was used to create a timed annotation for Guo Degang, Yu Qian, as well as the audience, which was considered as a single participating agent (Figure 2). These three primary annotation levels marked the respective start and end times for each phrasal utterance by the three performing agents, including a transcription of their dialog in Mandarin Chinese.
The subject video contained subtitles, but these were not always reflective of the actor’s actual speech; consequently, the transcription used in this study was provided by a native speaker of Chinese. To the phrasal annotations were added English glosses, and then detailed word-level annotations for the two performers, where each word in the transcription was aligned with the beginning and ending of its pronunciation in the dialog. The audience dialog was not broken up beyond a phrasal level, as their utterances, consisting of applause, laughter, etc., were not internally divisible.

These annotations were then imported into Praat for further analysis. Praat—developed by Paul Boersma and David Weenink of the University of Amsterdam—is a software suite for phonetic analysis widely used in linguistic research. Praat provides extensive tools for quantitative and statistical analysis of speech duration, quality, and intonation. These tools are accessible through the Praat GUI (Figure 3), as well as the Praat scripting language, which is useful for the automated analysis of large quantities of data.

Praat also supports the synchronization of audio with annotation files. Although it lacks the video capabilities or tier hierarchy of ELAN, it has more fine-tuned control over annotation timing and audio playback. The standard Praat sound viewer presents amplitude and spectrum graphs above the annotations. Other Praat tools can identify the pitch and intensity for a given interval in the recording, and these tools were used for the pitch matching analysis in this project.

After importing the ELAN transcriptions to Praat, we refined the annotation start and end times for all phrases and words, as Praat allowed for greater precision. Then we wrote Praat scripts that automated the recording of pitch and duration for all words in the two selected bouts, allowing for the analysis of pitch and beat matching between the two actors and the audience.

Analysis of Bout 1

The relationship between Guo Degang and Yu Qian in this piece is an example of what is called the yitouchen or “heavy-on-one-end” style of crosstalk, meaning that one actor—Guo Degang, in this case—is clearly dominant. In this style, the second actor (Yu) functions as a subsidiary performer, but his foil-like character is completely necessary to the success of the main actor (Lawson, 2017, pp. 91–94). In the same way that the interlocutor responds with a type of backchannel-like comment to the main speaker in a successful spoken bout in the research on pairs of same-sex friends, so this style of crosstalk can only be successful with the proper supportive co-narration provided by the straight man (Moser, 1990). The bout begins with Guo as the main speaker (in red), followed by Yu’s short phrases (in blue) that are less than one second to every 2–4 seconds of Guo’s speaking, seen in Figure 4.

Audience laughter also becomes increasingly prominent at this point in the bout, with the durations of audience responses eventually surpassing the durations of Yu’s utterances. The laughter increases as Guo becomes more
outrageous and Yu responds with greater incredulity. For example, when Guo muses about becoming 140 years old at the beginning of second 35 in Figure 5, Yu expresses disbelief with “Huh?”, triggering audience laughter, which Yu follows with “Good Lord” in seconds 37–38, provoking even more audience laughter.

Significantly, the dovetailing of responses by the audience, Yu, and Guo is clear from second 36 in the first line through the second line of Figure 5, with each agent building on another’s previous utterance and then setting up responses for the subsequent agent. The co-construction taking place between Guo, Yu, and the audience in this crosstalk aligns with the following observation by Kecske: communication is “a trial-and-error process that is co-constructed by the participants [involving] breakdowns, misunderstandings, struggles, and linguistic
aggression as properties which are in no way unique, but rather represent common features of communication” (2010, p. 70). Notably, in crosstalk, carefully scripted break-downs, misunderstandings, and even examples of linguistic aggression in the spoken banter between the two actors—aping real-life conversations—constitute the heart of the humor, creating the dynamic interplay between audience and actors throughout the performance. Clearly, crosstalk writers and actors understand the nuanced co-construction of everyday dialogue and exploit their expertise for maximum comedic benefit throughout the performance.

By second 40 the energy generated by the interactions among Guo, Yu and the audience has ramped up considerably since the beginning of the bout. The climax of the bout occurs when Guo claims that his acting partner, Yu, will still be performing with him at his advanced age. When Yu asks how he will still be there, Guo points to an urn on the stage—the one that will carry Yu’s cremated remains. At this point there is significant audience response as they laugh at Guo’s joke about Yu. Figure 6 shows how the audience’s loud response to Guo’s stated reason for the presence of the urn eclipses the utterances of both actors in duration and intensity, signaling the end of Bout 1.

Looking at Bout 1, we see the speech with backchannel-like communication or “heavy-on-one-end” mode of interaction between the main performer and his foil. As the bout progresses, Yu Qian’s comments get longer and more heated as he responds to Guo’s increasingly ridiculous statements, and the audience’s responses eventually exceed Yu’s in duration. The following analysis of the durations of

![Figure 5. Audience laughter (green) at Guo (red) in the top line contributes to Yu’s response (blue), which, in turn, contributes to more audience laughter at the beginning of the second line.](image)

![Figure 6. Audience laughter about the urn that will house Yu’s remains.](image)
utterances by Guo, Yu, and the audience demonstrates that the audience is clearly a major participant in this bout, albeit in a different way than Turino’s explanation of participatory interaction.

Analysis of Duration Data for Bout 1

We looked at the durations of utterances for each of the performing agents: Guo, Yu, and the aggregate response of the audience.

Data

The data consisted of ordered pairs of items collected in the order of the performance. The first item in each pair was a categorical indicator variable that indicated which agent (Guo, Yu, or Audience) made the utterance. The second item in the pair was the duration of each utterance measured in milliseconds.

Analysis

The analysis first consisted of an analysis of variance followed by Tukey’s HSD pairwise comparisons of the three agents. Second, a time series plot and a spectral analysis was conducted to see if there was any regular cyclical pattern in the duration data over the course of the performance.

Results

ANOVA

There were significant ($p = .0015$) differences in mean duration of the three agents, seen in Figure 7. The least squares means showed that the mean duration for Guo was the longest at 2044.6 milliseconds. The average audience utterance (1963.4 milliseconds) was second longest, and the average utterance for Yu (817.5 milliseconds) was the shortest. Tukey’s HSD test showed that there was no significant difference between the durations of Guo’s utterances and those of the audience ($\alpha = 0.05$ level of significance), but that the average duration of utterances by Yu was significantly shorter than that of either Guo or the audience (again at the $\alpha = .05$ level of significance).

Checking Assumptions of ANOVA

There was one extremely long utterance by Guo that did not match the pattern of the other data. However, including this outlier did not change the conclusions with respect to differences in the duration of utterances between agents. There was also a tendency for the variance in duration to increase as the average duration increased. However, performing a natural log transformation of the duration data to correct for this increase did not alter the conclusions.

Time Series Analysis

The time series plot and spectral analysis showed that there were no significant regular cycles in the duration data (Fisher’s Kappa = 4.11, $p = .309$).

Discussion of Duration Data

Significantly, although we did not find regular cycles in the duration data, we did find examples of periodicity when analyzing the speech stream for approximately 50% of the analyzed sections. The analysis of periodicity data follows the discussion of duration data in Bout 2.

Analysis of Pitch Data for Bout 1

We considered each juncture between the pitch at the end of an utterance by one agent (Guo, Yu, or audience) and the initial pitch of the following response by the subsequent agent, similar to the turn transitions studied by the Cambridge group as outlined in Robledo del Canto et al. (2016). We did notice an occasional matching of pitches—seen, for example, in Figure 5 (in seconds 36–40), in which the audience, Yu, and Guo built on one another’s responses after the joke about Yu’s ashes in the urn. However, when considering all of the pitch relationships between turns in the entire Bout 1, we did not identify any significant pitch matching.

Data

The data consisted of the ordered pairs of numbers in which the first number in the pair was the pitch of the last syllable of the first utterance by Guo, Yu, or the audience, and the second number was the starting pitch of the following response by Guo, Yu, or the audience. In addition, there was an indicator variable that distinguished whether the audience or one of the actors gave the first utterance and another indicator variable that indicated whether the audience or one of the actors gave the response. (Note: Source data table AudienceSpeakerAudienceSpokenPitch Data.jmp)
among the three agents, as illustrated in Figure 8. Moreover, after fitting the linear model it was determined that the linear correlation was not significantly \((p = .2436)\) different between the cases where the audience made the first utterance and the cases where one of the actors made the first utterance. Additionally, there was no significant \((p = .4707)\) difference in average pitch of the first syllable of the response between the cases where it was made by the audience or by one of the actors.

In sum, while the durations of audience utterances were significant enough to allow us to conclude that the audience was a participating agent in this bout, there was no significant \((p = .6048)\) linear correlation \((r = -.1232)\) between the pitch of the last syllable of the first utterance and the pitch of the first syllable of the subsequent response among the three agents, as illustrated in Figure 8.

Bout 2, however, provided very different results with regard to the pitch data and reinforced the conclusion about the significance of the durations of audience utterances.

**Analysis of Bout 2**

For the most amusing and animated part of the performance, Guo proposes that the two actors sing part of a difficult opera aria, with entreaties from Yu that the key not be too high. In this last part of the performance, Yu takes a more dominant role compared to the previous bout, illustrating a different kind of interaction that reflects the *zimugen* style or “two-sides-of-a-snap” mode—the Chinese description of a more equal exchange between the two actors (Lawson, 2017, pp. 92–95).

The second bout begins at 12:50 after Guo says, “There is a Beijing Opera called *Si Lang Visits his Mother*, and one of the arias has some fast rhythms and is especially high; let’s listen to Mr. Yu’s vocal rendition.” Then, enjoying the fact that Yu dreads singing after Guo’s overly generous introduction, Guo teases Yu by saying, “This is sure to be good, don’t worry. I’m here. And if you die, you will die quickly—no pain.” Looking at Figure 9, the audience begins laughing right as Guo says, “This is sure to be good” (beginning of line 1) and continues laughing during Yu’s very animated response of “Ai ya, Good Lord!” (middle of line 1) and beyond. While Yu and Guo’s responses dominate at the beginning of line 1, note that ELAN also detected audience laughter while they were speaking.

For the next 20 seconds, Guo and Yu banter back and forth about the key of the aria, ‘Jiao Xiao Fan.’ Yu begins this section with an animated request for not singing too high, and Guo assures Yu that he won’t. Instead, Guo begins the aria by singing too low in line 1 of Figure 10, provoking cheering applause from the audience during Guo’s solo and beyond. The audience laughter picks up in intensity after Guo says, “OK, take it from here” in line 2, and continues through most of line 3. It becomes apparent that the audience is not only responding to Guo’s solo, but is also setting up Guo’s and eventually Yu’s responses.

Yu is exasperated by Guo’s low solo (seconds 43–48, at the top of Figure 11), which is punctuated with audience laughter, further exacerbating Yu’s anguish.

Guo counters Yu’s complaints by singing the low part much higher. Figure 12 begins with the last part of the high solo, prompting audience laughter while Guo is still singing. The laughter continues for 2 seconds afterwards, and then for 5 more seconds after Guo says to Yu, “OK, now take it from here”—a cruel but hilarious taunt that provokes Yu to exasperation once again at the end of line 3 in seconds 81 and 82.

When Yu complains that even professionals are unable to sing it that high, Guo says that isn’t true, claiming that he can do it. Guo tells Yu to start with the initial low part of the phrase, which he will follow (seconds 92–100). Yu
doesn’t believe him, but nevertheless starts to sing his part anyway at the beginning of Figure 13. The audience is quiet as he sings and claps immediately afterward for less than one second before Guo ends the performances with an extremely high falsetto response to Yu in seconds 120 through 124 in the second line of Figure 13. The audience applauds with wild enthusiasm throughout Guo’s final seconds of the performance, with Yu matching Guo’s ending pitch in second 124 in the second line, completing the performance with his acknowledgement of Guo’s accomplishment, as seen in Figure 13.

Analysis of Duration Data for Bout 2
Looking at Bout 2, the sum of the durations of audience utterances in this bout is greater than the combined durations of the responses by both performers. Clearly, the total duration of the audience responses in Bout 2 demonstrates even more emphatically that the audience is a third performing agent in this crosstalk.

Data
The data consisted of ordered pairs of items collected in the order of the performance. The first item in each pair was a categorical indicator variable that specified which participant (Guo, Yu, or Audience) made the utterance. The second item in the pair was the duration of each response measured in milliseconds.

Analysis
The analysis first consisted of an analysis of variance followed by Tukey’s HSD pairwise comparisons of the three participants. Secondly, a time series plot and a spectral analysis was conducted to see if there was any regular cyclical pattern in the duration data over the course of the performance.

Results
ANOVA. There were significant ($P<.0001$) differences in mean duration of the three participants. The least squares
Figure 10. Audience laughter and calls during and after Guo’s low solo.

Figure 11. Yu’s animated complaints about Guo’s low solo, accompanied by audience laughter.
means shown in Figure 14 show that the mean duration for
the audience was the longest at 3801.7 milliseconds. The
average utterance by Guo (1091.0 milliseconds) was sec-
ond longest, and the average utterance by Yu (1022.2 milli-
seconds) was the shortest. Tukey’s HSD test showed that
there was no significant difference in the durations of the
utterances by Guo and Yu (at the $\alpha = 0.05$ level of sig-
ificance), but the average duration of utterances by the
audience was significantly longer (almost 4 times longer)
than either Guo or Yu (again at the $\alpha = 0.05$ level of sig-
ificance).

Checking Assumptions of ANOVA
The residual durations from the model were long tailed. How-
ever, performing a natural log transformation of the
duration data corrected this issue but did not alter the conclusions.

### Time Series Analysis

The time series plot and spectral analysis showed that there were no significant regular cycles in the duration data (Fisher’s Kappa = 3.74, \( p = .748 \)). Nevertheless, we determined a level of rhythmicity when we calculated the pulse in Mandarin between all syllables in a given utterance, after manually demarcating the boundaries between syllables, as evidenced in the next section.

### Analysis of Periodicity Data

#### Rhythmic Entrainment Between Performers and Between Performers and Audience.

We found that the crosstalk performance exhibited rhythmic entrainment between the two performers and between the performers and the audience in the two bouts we studied. While the concept of entrainment is often rather broadly defined in order to allow for a range of disciplinary applicability (Montague, 2019), we define entrainment as the case in which the first syllable of an utterance falls near or on beat with the rhythm established by the previous speaker’s/audience’s utterance. In order to quantify this entrainment, we developed a method to calculate the pulse of an utterance, finding a period in milliseconds to best fit the syllables of the utterance. For a given segment of dialog, \( f(T) \) gives a number between -1 and 1, which describes the fit of a period \( T \) in milliseconds between syllables; \( S \) is the number of syllables, \( n \) is a certain syllable of the utterance, \( t_1 \) is the timestamp of the first syllable, and \( t_n \) is the timestamp of syllable \( n \):

\[
    f(T) = \frac{1}{S} \text{Re} \left( \sum_{i=1}^{S} e^{2\pi i n t_i / T} \right)
\]

A period \( T \) that does not describe the utterance returns an \( f(T) \) close to 0; a period that does describe the utterance will give a high \( f(T) \) above 0.5, indicating that all syllables in the utterance fell close to the pulse described by period \( T \). A fit score of 0.5 indicates that all syllables occurred on average a 16.66% interval away from a beat of the pulse. A metronomic utterance perfectly described by \( T \) would have \( f(T) = 1 \) and \( f(T) = -1 \), and would actually indicate a rhythm that matches the utterance but is offset by half a beat. We used an algorithm to evaluate candidate values for \( T \), selecting the period with the highest value for \( f(T) \).

This method was successful for identifying a pulse for the utterances in the examined portion of the dialog. It identified a period with \( f(T) \) greater than 0.5 for 67 out of 72 utterances, indicating a high degree of fit. The median fit for all utterances was 0.73. Figure 15 shows an utterance with fit of 0.5, Figure 16 with the median fit 0.73, and Figure 17 shows a high fit of 0.95. The top row of text is performer Yu Qian’s comments in Chinese, and the bottom shows Guo Degang’s comments. The black tick marks indicate the midpoint of the syllable, and the red tick marks

---

**Figure 14.** Least squares means and results of Tukey’s HSD test.
demarcate the best fit rhythm. The average period for all utterances was 181 ms, with values ranging between 100 and 200 ms. Both performers varied within this range, and there was not an overall difference in the rhythms of the two performers.

According to Ding et al. (2017), the modulation rate of speech across a range of languages, including English and Chinese, is 5 Hz, or approximately 200 ms; for music, the authors localize the rate to 2 Hz, or about 500 ms. London (2012, p. 27) asserts that “the shortest interval that we can hear or perform as an element of rhythmic figure is...about 100 milliseconds,” making the average of 181 ms in the present study somewhat lower than the average modulation rate calculated by Ding et al. (M. Wilson & T. P. Wilson, 2005). We believe the reason for these slightly lower values is that crosstalk is a kind of performed speech in which semantic comprehension is paramount. As a consequence, Chinese performers in the narrative arts traditions are trained to enunciate clearly and probably speak more slowly and deliberately than in normal parlance in order to make sure meaning is communicated unambiguously (Lawson, 2011, pp. 83, 93). Thus, in the same way that intelligibility may override periodicity in performance (Hawkins, 2014), so semantic comprehension may override faster modulation rates compared to normal speech.

Rate of Performer-Performern Entainment. We describe entrainment between performers as occurring when the first syllable of one performer arrives on beat with the pulse established by the previous turn of the other speaker. The same fit function described above is used to evaluate its fit, where it simplifies to the following form; the summation is not needed as only one point is considered. The value of $t_1$ is the time of the first syllable of the first turn, which established the period $T$, and $t_2$ is the first syllable of the following turn, being measured for entrainment.

$$f(T) = \text{Re} \left( e^{\frac{2\pi i (t_2-t_1)}{T}} \right)$$

This formula calculates how closely the measured syllable aligns with the period $T$ defined by the previous turn, and gives a value between $-1$ and 1 that correlates to the fit equation given above.

We considered a turn to be entrained to the previous turn when its first syllable landed within 16.66% of the pulse period, as indicated by a fit score greater than 0.5. Of the 72 turn-response pairs analyzed, 35 passed this threshold, showing that half of the pairings in the studied portion of the performance exhibited rhythmic entrainment. Furthermore, 12 of the utterances exhibited off-beat, syncopated entrainment, with their first syllable falling within 15% of the exact offbeat established by the previous turn. The remaining 15 turns showed no entrainment of any kind. Figure 18 shows an example of entrainment, as the first syllable of the second utterance spoken by Yu Qian falls exactly on the pulse established by Guo Degang’s preceding turn.

Rate of Performer-Audience Entrainment. Finally, we examined entrainment of the audience responses to the performers. Although audience utterances consisted of general periods of noise and could not be divided into syllables, we did identify the moment of onset for 26 audience reactions. This was done by manually marking the moment when the audience is first heard in the recording for a given utterance. These moments of onset were evaluated in the same manner as the first syllable in inter-performer turns, as described in the previous section; the onset of audience reaction was evaluated against the pulse established in the most recent performer turn.
Similar to the performer-performer data, the audience was rhythmically entrained to the performer pulse half of the time. Out of 26 measured reactions, 14 scored higher than 0.5 in the fit equation; in fact, these 14 entrained reactions all scored higher than 0.8, indicating that these moments of onset occurred within 10% of the pulse interval, a high degree of rhythmic alignment. Similarly, 7 onsets occurred within 15% of the offbeat, with 5 onsets that showed no relation to the established pulse.

**Discussion of Pulse Data.** Approximately 50% of the analyzed sections from the crosstalk performance exhibited rhythmic entrainment between the two performers and between the performers and the audience members.

**Analysis of Pitch Data for Bout 2**

In addition to the periodicity found in approximately 50% of the analyzed sections in Bouts 1 or 2, there was a highly significant ($P<.0001$) linear correlation ($r = .6947$) between the pitch (in Hz) of the last syllable of the first utterance and the pitch of the first syllable of the response, illustrated in Figure 19. The results indicate that the first pitch of the response tends to be similar to the pitch of the last syllable in the first utterance. Further, after fitting the linear model, it was determined that the linear correlation was not significantly ($p = .8691$) different between the cases where the audience made the first utterance and the cases where one of the actors made the first utterance. Additionally, there was no significant ($p = .1375$) difference in the average pitch of the first syllable of the response whether it was given by the audience or by one of the actors.

In other words, in Bout 2, Guo, Yu, and the audience members attempted to approximate the pitch of their response to the ending pitch of the preceding agent’s utterance. While Guo and Yu might have been expected to match pitches since the point of the bout was for one of them to start and for the second one to finish a difficult opera aria, the pitch-matching approximations did not occur solely during the sung portions of the bout. Moreover, the audience—in aggregate—attempted to approximate pitches with the performers, who, in turn, also attempted to match the pitches of the preceding audience utterances. The pitch matching occurred spontaneously between performers and audience members throughout the final bout.

The anticipatory and improvisational interactions between actors and audience members through reciprocal pitch matching and long audience responses fueled the energy throughout the performance. Tsioulakakis (2013, p. 6) has proposed the term “performative mutuality” as an optimal experience of live music making, and we suggest that performative mutuality is also an appropriate term to describe performer-audience interaction in this crosstalk.
Conclusion

Results from this project suggest that when interactants are attitudinally aligned in this crosstalk performance, speech may become (1) rhythmically entrained for part of the performance and (2) musicalized in terms of pitch approximation through the process of co-narration by the end of the performance. While the data from the Cambridge study underscores how both pitch and pulse are operationalized in measuring the speech stream across turns in the successful bouts between same-sex friends, the data from the crosstalk performance suggests an underlying periodicity for approximately 50% of the studied segments of the performance and a gradual pitch-alignment phenomenon across the turns among the three performing agents during the final bout of the performance. The most probable reason for the 50% rate of periodicity in crosstalk is that the presentational speech of crosstalk is the focal point of the performance, thereby overriding the rhythmicity that can occur in the successful bouts between same-sex speakers (see Hawkins, 2014, p. 5). Thus, we argue that the importance of conveying the communicative function of language is one of the most important reasons for the lack of rhythmicity in 50% of the analyzed samples in the crosstalk, as well as the reason for the generally slower modulation rate.

Additionally, the presentational nature of a crosstalk performance constrains the audience reaction, thereby reducing regular shared attention and periodicity throughout the crosstalk. If rhythmic entrainment between performers and audience members were constant through a performance, it would be too predictable and, therefore, uninteresting to audiences who pay to be entertained by the comedic dialog of professional performers. Nevertheless, the speech between interactants in crosstalk became progressively more musicalized by the end of the performance, demonstrating a strong affiliative connection through the charged emotional bonds created by the comedic performance and the audience’s enthusiastic responses. While the results from this study suggest a relationship between strong affiliative interactions between the three agents of performance through rhythmic entrainment for about half of the performance and increased pitch alignment in their utterances at the conclusion of the performance, more research is needed in order to begin to generalize about rhythmic entrainment and pitch matching across turns in comedic performance.

The attitudinal alignment and increased pitch approximations across turns in this crosstalk was initially catalyzed by the skills of the writers and actors. By simulating the natural way two speakers communicate, crosstalk writers and actors combined their skills to create a realistic-sounding comedic co-narrative as a presentational performance. After years of performing crosstalk together, Guo and Yu have become virtuosi of the genre, mastering the rhythms inherent in their back-and-forth repartee. But crosstalk is far more complicated than merely witnessing two actors co-constructing dialogue on stage. When factoring audience interaction into a crosstalk performance, one sees an additional dimension of performative mutuality or cooperative co-narration involving the audience vis-à-vis the actors. At the same time the actors engaged in comedic banter, they also beckoned the audience to participate in the performance through laughter and cheering applause—responses that happened spontaneously as the audience community reacted, as a single body, to the antics of the performers on stage.

In many ways, the behavior of the crosstalk audience aligns with what Himberg (2013, p. 38) says about the behavior of musical participants:

> When people make music together, they coordinate their behaviors in time. Typically, the result is that a regular pulse structure emerges that is experienced as shared, around which each participant organizes their contribution to the musical event…participants entrain their attention, actions and sounds with those of the other participants, mutually adapting their behaviors in time (Cross, 2014, p. 813).

Although crosstalk is technically not a musical performance, nor is it participatory in the way Turino describes it, the audience certainly entrained their attention and sounds with one another as they reacted en masse to the performers, and the actors similarly entrained their performance rhythmically and melodically with the utterances and mindset of the audience. If it had been possible to document the actions of the audience members, one might have also seen entrained facial expressions and bodily movements as well.

Significantly, the melodic entrainment between actors and audience members also increased during the course of the performance. During the first bout, the audience was involved vocally, but not nearly as much as they were during the final bout. Leading up to the baofu, the increased pitch approximation during turns corresponded to an increased intensity in the affiliative and emotional connection between the actors and their audience. Thus, as the bonds between audience and performers became more intimate and emotionally charged, the melodic aspects of speech increased, particularly at the climax of the piece.

Put another way, the performed speech of the actors at the beginning of the performance eventually changed from a presentational to a participatory mode as Yu and the audience members became gradually more involved. Based on the changing modes of interaction in this crosstalk, we argue that presentational and participatory interactions are not mutually exclusive, but rather complementary and com-mutable types of performance. “Presentational” might simply be seen as the heavy-on-one-end mode of interaction in which the secondary speaker—or the audience—begins with a more subsidiary albeit still minimally participatory role. On the other hand, “participatory” interaction might
simply be the two-sides-of-a-snap mode of performance, in which the speakers—or speakers and audience members—share the conversation more equally. Thus, the movement towards pitch-aligning speech co-occurred with an increased sense of participation among all performing agents, resulting in mutual bonding between actors and audience members over the course of the performance.

If music has an advantage over language in being able to integrate the experience of multiple participants and create a collective communicative interaction (Cross et al., 2016, p. 66), then we assert that the pitch-aligning speech in crosstalk does essentially the same thing by integrating audience members into a community of participants who function essentially as a third performing agent alongside the two actors on stage. Rather than thinking of speech and music as separate domains, we agree with the researchers of the Cambridge study that it is better to construe them as overlapping categories of interactive, communicative behavior. Moreover, we suggest that the performative mutuality demonstrated in the crosstalk performance exhibits interactivity in the complementary presentational and participatory modes of performance as well as in the interdependent and mutually reinforcing realms of music and language.

**Acknowledgement**

We gratefully acknowledge the suggestions and comments made by the two reviewers. The paper is much improved because of their careful readings.

**Contributorship**

Francesca R. Sborgi Lawson conceived of the paper, and Francesca and Joshua Sims designed the study. Joshua Sims analyzed the data in ELAN and Praat, and John Lawson analyzed the data in R. All the authors reviewed, edited, and approved the manuscript.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Francesca R. Sborgi Lawson’s research fund, supported generously by the College of Humanities at Brigham Young University through her professorship, paid for Joshua Sims to analyze the data in ELAN and Praat.

**ORCID iD**

Francesca R. Sborgi Lawson https://orcid.org/0000-0002-5571-3214

**Action editor**

Elizabeth Tolbert, Johns Hopkins University, The Peabody Institute.

**Peer review**

Youn Kim, University of Hong Kong, Department of Music.

Ian Cross, University of Cambridge, Faculty of Music.

**Supplemental material**

Supplemental material for this article is available online.

**Notes**

1. See Lawson (2011, pp. 115–120), Tsau (1979-80), Link (1980, 1986), and Moser (1990) for an introduction to the genre, and Cai (2016, pp. 83–87) for an introduction to Guo Degang, the main performer featured in this analysis.

2. A complete side-by-side translation of the text of the performance from Chinese to English is included in Appendix A. The authors wish to acknowledge the contribution of Joseph Qiwei Zhang in transcribing and editing the Chinese text of the crosstalk and in editing the English version of the text, translated by F. Lawson.

3. SMG上海电视台官方微博 SMG Shanghai TV Official Channel. (2016, February 15).

   Guoode Gang Yaqian Cross Talk 笑傲江湖 2016 SMG Spring Festival [SMG Official HD] [Video file]. Retrieved from https://www.youtube.com/watch?v=kobiKhm5g_o

4. The phenomenon of the audience performing as a collective body is discussed in greater detail later in the article.

5. Since the video-recorded performance only occasionally spotlights the faces and upper-body movements of a few of the 500+ audience members, we concentrated on the utterances made by the audience and the two performers rather than on the visual and kinesthetic aspects of the performance.

6. The phenomenon in which the audience functioned as a single entity in the analysis is discussed in greater detail later in the article.

7. See Note 2.

**References**

Cai, S. (2016). A culture hero: Xiangsheng (crosstalk) performer Guo Degang. *Asian Theatre Journal*, 33(1), 82–103.

Chen, J. Y., Lin, W. C., & Ferrand, L. (2003). Masked priming of the syllable in Mandarin Chinese syllable production. *Chinese Journal of Psychology*, 45, 107–120.

Chen, R. (1993). Responding to compliments: A contrastive study of politeness strategies between American English and Chinese speakers. *Journal of Pragmatics*, 20, 49–75.

Chen, T. M., & Chen, J. Y. (2012). The syllable as the proximate unit in Mandarin Chinese word production: An intrinsic or accidental property of the production system? *Psychonomic Bulletin and Review*, 20, 154–162.

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research*, 57(4), 1468–1479.

Clayton, M. (2007). Time, gesture and attention in a Khyal performance. *Asian Music*, 38(2), 71–96.
Clayton, M., & Leante, L. (2015). Role, status and hierarchy in the performance of North Indian classical music. *Ethnomusicology Forum*, 24(3), 414–442.

Cross, I. (2014). Music and communication in music psychology. *Psychology of Music*, 42(6), 809–819.

Cross, I., Hawkins, S., Ogden, R., Robledo, J. P., Gill, S., London, J., Ross, B., Liu, J., Knight, S., & Halford, D. (2016). *Music as communication, communication as music*. Paper presented at the annual meeting of the Society for Ethnomusicology, November 2016, Washington, DC.

Gorisch, J., Wells, B., & Brown, G. (2012). Pitch contour matching and interactional alignment across turns: An acoustic investigation. *Language and Speech*, 55(1), 57–76.

Hawkins, S. (2014). Situational influences on rhythmicity in speech, music, and their interaction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1658), 20130398.

Hawkins, S., Cross, I., & Ogden, R. (2013). Communicative interaction in spontaneous music and speech. In M. Orwin, C. Howes, & R. Kempson (Eds.), *Music, language and interaction* (pp. 285–329). College Publications.

Himberg, J. (2013). *Interaction in musical time* (Ph.D. Thesis). University of Cambridge.

Keil, C. (1987). Participatory discrepancies and the power of music. *Cultural Anthropology*, 2(3), 275–283.

Kecskes, I. (2010). The paradox of communication: Socio-cognitive approach to pragmatics. *Pragmatics and Society*, 1(1), 50–73.

Lawson, F. R. S. (2011). *The narrative arts of Tianjin: Between music and language*. Ashgate Publishing.

Lawson, F. R. S. (2017). *The women of Qiyi: Liminal voices and androgynous bodies*. Routledge.

Leante, L. (2017). Observing musicians/audience interaction in North Indian classical music performance. In I. Tsoumakis & E. Hytonen-Ng (Eds.), *Musicians and their audiences: Performance, speech and mediation* (pp. 34–39). Routledge.

Levow, G. (2019). *Turn-taking in Mandarin dialogue: Interactions of tone and intonation* (Unpublished paper). Computer Science Department, University of Chicago.

Li, S. (1985). *A brief study of the writing of traditional Xiang Sheng texts* (Unpublished paper).

Link, P. (1980). Hou Baolin: An appreciation. *Chinese Literature*, 2, 84–94.

Link, P. (1986). Stuck in Xiangsheng. *Chinoperl Papers*, 14, 27–36.

Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2), 179–214.

London, J. (2012). *Hearing in time: Psychological aspects of musical metre* (2nd ed.). Oxford University Press.

Montague, E. (2019). Entrainment and embodiment in musical performance. In Y. Kim & S. L. Gilman (Eds), *The Oxford handbook of music and the body* (pp. 177–192). Oxford University Press.

Moser, D. (1990). Reflexivity in the humor of Xiangsheng. *Chinoperl Papers*, 15, 45–68.

Nissen, S. (2019, August 6). Phone interview with author. Department of Communications Disorders. Brigham Young University. Provo, UT, USA.

Ogden, R., & Hawkins, S. (2015, 10-14 August). Entrainment as a basis for coordinated actions in speech. In *18th International Congress of Phonetic Sciences*, Glasgow. https://internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/proceedings.html

Roach, P. (1982). On the distinction between stressed-timed languages and syllable-timed languages. In D. Crystal (Ed.), *Linguistic controversies: Essays in Honour of F. R. Palmer* (pp. 73–79). Arnold.

Robledo, J., Hawkins, S., Cross, I., & Ogden, R. (2016). Pitch-interval analysis of ‘periodic’ and ‘aperiodic’ Question + Answer pairs. In H. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Speech prosody* (pp. 1071–1075).

Shen, X. S. (1993). Relative duration as a perceptual cue to stress in Mandarin. *Language and Speech*, 36(4), 415–433.

Tsau, S. (1979-80). Xiangsheng and its star performer, Hou Baolin. *Chinoperl Papers*, 9, 32–79.

Tseng, C. Y., & Lee, Y. I. (2004). Speech rate and prosody units: Evidence of interaction from Mandarin Chinese. Paper presented at Speech Prosody 2004. http://www.isca-speech.org/archive/sp2004

Tsoulakas, I. (2013). The quality of mutuality: Jazz musicians in the Athenian popular music industry. In F. Holt & C. Wergin (Eds.), *Musical performance and the changing city: Post-industrial contexts in Europe and the United States* (pp. 200–224). Routledge.

Turino, T. (2008). *Music as social life: The politics of participation*. University of Chicago Press.

van Gulik, R. H. (1969). *The lore of the Chinese lute: An essay in Ch'in Ideology*. Monumenta Nipponica Monographs.

Wilson, M., & Wilson, T. P. (2005). An Oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12(6), 957–968.

Yung, B. (2010). Music of *Qín*: From the scholar’s study to the concert stage. In J. H. Y. Cheung & K. C. Wong (Eds.), *Reading Chinese music and beyond* (pp. 9–28). City University of Hong Kong Press.