Author(s): Fischinger, Timo; Frieler, Klaus; Louhivuori, Jukka

Title: Influence of virtual room acoustics on choir singing

Year: 2015

Version:

Please cite the original version:

Fischinger, T., Frieler, K., & Louhivuori, J. (2015). Influence of virtual room acoustics on choir singing. Psychomusicology: Music, Mind, and Brain, 25(3), 208-218. https://doi.org/10.1037/pmu0000117

Influence of Room Acoustics on Choir Singing

Timo Fischinger[a], Department of Music, Max Planck Institute for Empirical Aesthetics,
Grüneburgweg 14, 60322 Frankfurt am Main, Germany

Klaus Frieler, Institut für Musikwissenschaft Weimar-Jena, Hochschule für Musik Franz Liszt,
Platz der Demokratie 2/3, 99423 Weimar, Germany

Jukka Louhivuori, Department of Music, University of Jyväskylä, Seminaarinkatu 15, P.O. Box
35, FI-40014 Jyväskylä, Finland

Running title: Influence of Room Acoustics on Choir Singing

Date on which the manuscript was uploaded to the system: June 8, 2015

a) Author to whom correspondence should be addressed. Electronic mail:
timo.fischinger@aesthetics.mpg.de

Abstract

Multitrack recordings of a mixed adult choir with 23 singers were collected in order to investigate the influence of varied room acoustical conditions on a choir's performance with regard to intonation, loudness, tempo, and timing precision. Headset microphones were used to record each chorister separately while the collected sound of all singers was presented via headphones exerting acoustic simulations of rooms with different reverberation times of 0.0 (bypass), 1.77 and 4.79 s according to three singing conditions. The choir was asked to sing "Locus Iste" by Anton Bruckner (1824-1896). Objective measures were obtained from single audio track analyses using the monophonic pitch tracker pYIN plugin for Sonic Visualiser. These revealed that intonation was barely affected by simulated room acoustics whereas tempo was notably slower and timing precision declined in the condition where participants sang in a comparatively large virtual room. Subjective judgments gathered by a questionnaire inquiring on the singers' experiences showed a clear preference for singing in the medium sized "Concertgebouw" (with reverberation time = 1.77 s), while the dry acoustical condition (bypass) was felt to be the best to sing in time. The significance of these results and their relationships to other musical and acoustical parameters are discussed.

PACS number(s): 43.75.Rs, 43.75.Yy, 43.75.Bc, 43.75Cd

I. INTRODUCTION

The aesthetic appreciation of a choir performance heavily relies on both the singers' skills and the acoustical characteristics of the venue. Choir directors usually know that choral performances are greatly influenced by room acoustics, while the choir singers experience the difference between singing in a small room for practice and performing in a comparatively large space like a concert hall.

Investigations into the interplay of acoustics and architecture in ancient Greek and Roman theaters reveal that architects from this early period, such as Vitruvius (first century B.C.), were already aware of the physical aspects of sound wave propagation and general aspects of room acoustics (Declercq & Dekeyser, 2007; Farnetani, Prodi, & Pompoli, 2008). In later times, scholars like Athanasius Kircher (1602-1680) systematically explored the characteristics of acoustic spaces through various experiments on the reflection of sound, as documented in part IV of the ninth book of his Musurgia universalis (1650).

Likewise, early composers were knowledgeable about the acoustical features of the locations in which their music was performed (in churches, concert halls, chambers, open-air, etc.). Indeed, the room size of a venue was reflected in certain composition rules, the way of instrumentation, and specific styles of performance of that time. For instance, in contrast to traditional (unison) Gregorian chants, melodic lines (with separate voices) in Renaissance polyphony had to be composed in a specific manner to avoid stylistically unsuitable dissonances that could occur due to late early reflections and very long reverberation times in large rooms. Additionally, composers of the late Renaissance and early Baroque period developed the Venetian polychoral style by placing groups of singers at different positions in a church in order to adapt to the acoustical conditions of (large) churches in Venice (e.g., Zarlino, 1558). Though this kind of musical conceptualization remains a special case of interaction between music and architecture, it shows that composers had a considerable knowledge of the influence of room acoustics. Later on, musicians and composers like Quantz (1752) and Mozart (1756) provided a number of recommendations regarding room size and how to adjust the style of performance (e.g., tempo adjustments) to the acoustical characteristics of a venue. For instance, Quantz (1752) recommends playing slower in large rooms compared to playing in small chambers to preserve the intelligibility of the music.

Taking into account the important role of acoustics for music performances, it is surprising that not much research has been done in this area from the musicians' point of view as compared to the listener's perspective which is a crucial factor in designing concert halls. Clearly, it would be beneficial for musicians to understand the effect of room acoustical features on their performance and how best to adjust tempo, phrasing, dynamics, and other musical parameters with respect to a given venue's acoustical environment.

Empirical studies on the influence of room acoustics on solo music performance revealed that up to 50% of a performance feature's variance such as e.g. tempo or loudness may be explained by room acoustical parameters (Schärer Kalkandjiev & Weinzierl, 2013). Solo musicians seem to intuitively adjust their performance to the room's acoustical situation, with "tempo" being one of the parameters significantly influenced by the specific reverberation time (RT) of the music venue (Schärer Kalkandjiev & Weinzierl, 2015; Ueno, Kato, & Kawai, 2010). According to these findings, musicians tend to play slower in rooms with very long and very short RTs.

In a similar way, individual singer's intonation and timing might be influenced by the level at which singers hear their own voice and the rest of the choir due to the characteristics of acoustical feedback (Sundberg, 1987; Ternström, 1989; Ternström & Sundberg, 1988).

To evaluate the effects of room acoustics among other relevant aspects like musical material, type of choir and vocal effort, Ternström (1993) recorded three choirs (a boys' choir, a youth choir and an adult choir) singing in different rooms (a rehearsal hall, a basement room and a large church) using two microphones for each recording session. The analyses of his recordings mainly revealed a large effect on the shape of the Long-time Average Spectra (LTAS) caused by the room acoustics, whereby choirs seemed to adapt their sound level and general usage of their voices to the acoustical characteristics of the venue, e.g., reflections of the room. A study on the directivity and auditory impressions of singers conducted by Marshall and Meyer (1985) showed that ensemble singers commonly prefer strong early reflections, whereas loudness of the so-called reverberation sound (Sundberg, 1987) becomes more relevant if the distance to the nearest sound reflector exceeds 7 m.

Because different rooms vary in the amount of direct sound in late and early reflections as well as in reverberations from the diffuse field, it is important to carefully adjust the spacing between singers to find an appropriate formation for a choir (Daugherty, 2003). This is to ensure that the balance of loudness between each singer's own voice ("Self") and the sound pressure level from the rest of the choir ("Other") is comfortable and suitable for all singers. Using binaural microphones placed at the outer ears of each singer, Ternström (1994) was able to show that the average Self-to-Other Ratio (SOR) in live performance of a chamber choir (of 25 singers) is typically about +4 dB, i.e., singers usually like to hear their own voice a little bit louder than the sound of the other voices (the "reference"). In an additional laboratory study with choir singers, Ternström (1999) was further able to show that the average SOR value is about +6 dB (ranging between 0 and +15 dB), if the singers get the chance to control their preferred SOR individually when presented a synthesized choir over headphones as reference. This study also revealed that singers are extremely precise in the reproduction of their preferred SOR within a very low tolerance range of +/-2 dB. The average SOR values may be even higher, if measured from singers distributed over an opera stage (Ternström, Cabrera, & Davis, 2005). Nevertheless, different room absorption may also evoke differences in "vocal effort" and therefore lead to different intensity levels within a choir. Either way, particularly amateur singers vary to a high degree with regard to the strength of their voice and their dynamic intensity variations respectively (Coleman, 1994).

In order to illuminate both the inherent processes of choir singing in general and the underlying interactions concerning loudness balance in particular, it is necessary to focus on the singing behavior of each chorister separately. Accordingly, Jers and Ternström (2005) used multitrack recordings to investigate the differences in intonation quality between a professional and a semi-professional choir respectively. Although they did not find any statistically significant differences between the two ensembles, this method appears to be very promising. It offers insights into the complex multi-layered facets and interactions of choir singing like, for instance, the so-called "chorus effect" (Jers & Ternström, 2005). The chorus effect originates from the quasi-random and highly complex sound produced by the merging of many voices including their reflections.

Fischinger and Hemming (2011) were able to replicate the results of Jers and Ternström (2005) with regard to the mean fundamental frequency (MF0) and the corresponding standard deviation (SF0) using a multitrack recording setup similar to that of Jers and Ternström (2005).

Other important facets of choir singing concern the actual tuning of a performance as well as intonation drift or pitch drift (Howard, 2007b; Seaton, Sharp, & Pim, 2014). Studies on intonation within unaccompanied singing ensembles revealed that singers tend to prefer to sing in just intonation (Howard, 2007a). However, adapting their intonation on consonances to comply with non-equal tempered tuning systems, might result in a pitch drift over a whole piece of music

(Howard, 2007b) due to small incommensurabilities. A more recent study on intonation and intonation drift in unaccompanied solo singing by Mauch, Frieler, and Dixon (2014) observed in single cases a median absolute pitch drift of 11 cents over a duration of about 50 s. Contrary to Howard (2007a), Mauch et al. (2014) could not find any preference of solo singers for singing in equal temperament or just intonation.

The ability to sing in tune and with high precision and accuracy depends on the level of singing expertise or experience (Dalla Bella, Giguère, & Peretz, 2007; Pfordresher, Brown, Meier, Belyk, & Liotti, 2010). Dalla Bella et al. (2007) were able to show that pitch stability between repeated sequences of notes was less consistent and showed larger deviations in occasional singers (0.6 semitones) than in professional singers (0.3 semitones).

If asked to adjust their sung pitch in response to pitch changes of an external musical interval, highly skilled choir singers react slower (after 227 ms) compared to moderately skilled choir singers (after 206 ms) (Grell, Sundberg, Ternström, Ptok, & Altenmüller, 2009). This may be due to different procedures of action-perception (voice) control. Nevertheless, this study demonstrates how rapidly choir singers can adjust their individual pitch to a changing external pitch reference.

A relatively large number of studies on choir acoustics investigating singing behavior using individual measurements have focused on short musical excerpts. Only little is known about how choral performances are influenced by room acoustics (Ternström, 2003; Ternström, Jers, & Nix, 2012). Moreover, to our knowledge, there has been no attempt to investigate the influence of different room acoustics with varying RTs on choir singing using multitrack recordings under systematically varied and controlled feedback conditions. Thus, we tried to tackle the question of how the ease of intonation and (synchronization) timing as well as tempo, pitch drift, tuning, and loudness are affected by room acoustics. The goal was to use an ecological approach with a setup very similar to everyday choir singing practice.

Therefore, choir singers performing under three different virtual room acoustical conditions were recorded individually using multitrack techniques. Objective acoustic analyses as well as subjective measurements using a questionnaire on the judgments of the choir singers were then employed to investigate the influence of room acoustics on intonation, loudness, tempo, and timing precision.

II. METHOD

A. Participants

A mixed adult choir from Jyväskylä (Finland) with 23 singers (5 sopranos, 9 altos, 7 tenors, 3 basses) with a mean age of 45 years participated in the study. The average years of choir singing experience was 29 years and all singers reported normal hearing as well as no vocal pathology. The choir can be classified as an experienced choir with concerts in various countries and 13 CD recordings.

B. Materials and apparatus

Choir recordings were collected in a professional recording studio at the Department of Music at the University of Jyväskylä. For an appropriate formation of the choir, the singers were placed into two rows with spacing of 60-80 cm between individuals (side by side) and a distance of 1 m between rows. Through half-open headphones (AKG K-141 MK II) each singer heard all of the other singers (artificial airborne sound reference) as well as their own voice (artificial airborne sound feedback) as recorded by each headset microphone (AKG C 420 PP) and mixed by the studio mixer (AVID ICON D-Command with Pro Tools). The balance of loudness between each singer’s own voice (feedback) and the sound of the other singers of the choir (reference) was adjusted individually for each participant’s preferred SOR respectively. Important to note is that the singers also heard the bone-conducted sound of their own voice as well as the other singer’s voices from outside their headphones. The impression of this setup may have been a little bit different to normal singing without headphones, but the participants reported that they perceived this acoustical setup as being quite natural. Therefore the recording session can be considered as having high ecological validity. However, similar to the experimental setup of Ternström and Sundberg (1988), it was not possible to determine the precise SPL presented over the headphones. Research on bone-conducted and airborne sounds in speech by Pörschmann (2000) showed that both sounds are about equally loud. It can be assumed that this balance is similar for singing. The conductor of the choir was also equipped with headphones presenting the sum of all voices.

In order to evaluate the influence of varying room sizes (including different pre delays and RTs)[1] two different acoustics were selected from the Pro Tools reverb plugin TL Space: ”Concertgebouw” (AC2, RT = 1.77 s), and ”Spanish Cathedral” (AC3, RT = 4.79 s). The room acoustical simulations were based on stereo room impulse responses of the original venues. The setup was complemented with a bypass condition without any virtual acoustics added (AC1).

The choir was asked to sing “Locus Iste” by Anton Bruckner (1824-1896). The duration of this motet is around 3 minutes (medium slow tempo, ~80 bpm) and it contains strong dynamical changes from pianissimo (pp) to fortissimo (ff). The large range (ambit) requires the use of the whole register in each voice group, with a few modulations and a couple of demanding harmonies and unusual or large intervals (see score in the supplementary online section).

C. Procedure

The recordings lasted 90 minutes including warm-up, instructions and setup of headphones and headset microphones. The formation of the choir singers including the position of the conductor remained the same during the entire recording session.

After the setup of the experiment, the choir was asked to sing “Locus Iste” for three times under varied acoustical feedback conditions featuring three different virtual room sizes (VRS). The conductor and the singers were given the chance to get a short impression of each acoustical condition at the beginning of each recording, when humming the first notes of the score before the choir started to sing.

In order to collect subjective judgments, participants were asked to fill out a short questionnaire instantly after each recording. The questionnaire included six items related to their opinions/feelings about the acoustical condition during singing: “It was easy to sing”, “It was easy to sing in tune”, “It was easy to hear the voices of the other singers”, “It was easy to hear my own voice,”, “It was easy to sing in time”, “I was encouraged to sing”.

A five point Likert scale of agreement was used for each item (1 = totally disagree, 2 = somewhat disagree, 3 = neither disagree nor agree, 4 = somewhat agree, 5 = totally agree).

D. Data analysis

The experiment resulted in 23 voice tracks per condition, giving 69 tracks in total. (One singer did not want his recordings to be used.) All single tracks were analyzed using the pYIN plugin for Sonic Visualiser (Mauch & Dixon, 2014), which is one of the best monophonic pitch trackers currently available (Molina, Tardón, Barbancho, & Barbancho, 2014).[2]

However, the resulting pitch annotations still needed extensive manually corrections. Short events and other artefacts had to be removed; a few octave errors were transposed. Occasionally, sliding into the pitch by the singers resulted in two or more annotated pitches for one musical tone. In this case the tone were fused to one event by using the pitch from the steady-state phase and the onset (time) of the beginning slide-in part. Subsequently, all corrected pitch events were manually labeled with the corresponding note number of the “Locus Iste” score, and imported into the statistical software package R for further analysis. Hereby the original frequencies were converted to fractional MIDI pitch numbers based on concert pitch a=440 Hz. The final dataset had some peculiarities, since the pitch tracks typically do not contain annotations for every nominal note in the score, typically due to tone repetitions sung in legato for which the algorithm is often not able to find note boundaries. Consequently, the pitch annotations are slightly different for all singers and conditions with respect to total counts and notes annotated, but for all singers and conditions a sufficiently large set of pitch annotations of about 79-119 notes were collected.

The “Locus Iste” ground truth (bass = 94, tenor = 118, alto = 115, and soprano = 113 notes) was also manually coded and imported into R, where the metrical positions of the notes in the score were encoded by enumerating all possible 16th note positions.

III. RESULTS

A. Tuning and drift

For most intonation measures, except consistency, the reference to an external target pitch (or an interval derived therefrom) is needed, which was not given in the present experimental setup. However, consistencies alone are not sufficient to fully assess musical intonation (e.g., imagine a voice groups singing consistently one semitone higher or lower than the notated score pitch, which will result in high consistency but completely wrong intonation altogether).

Hence, some preliminary steps and checks had to be carried out to be able to use measures using target pitches. For example, it is not *a priori* clear if the singers use a particular tuning system, e.g. Equal Temperament, Just Intonation or Pythagorean tuning (Howard, 2007a). Furthermore, even if a tuning system is consistently employed by a choir there still could be an overall drift (Howard, 2007b; Mauch et al., 2014), i.e., a shifting of the reference pitch of the tuning system.

To this end, we first looked for possible drifts in the three conditions by regression on differences of sung pitch to a nominal pitch (equal temperament was chosen as an arbitrary reference, because the choice of a particular tuning system does not matter for measuring drift). We performed rank correlations of pitch differences with normalized onset (cf. below), and found a significant drift only in condition AC2 (Spearman's $\rho = 0.08$, $p<0.001$, other conditions: AC1: $p=0.898$; AC3: $p = 0.209$). However, the drift is only -4.6 cents along the whole course of the musical piece, which we deemed negligible.

Second, we checked if the pitches of all singers in all conditions do better fit to Equal Temperament (ET), Just Intonation (JI) or Pythagorean tuning (PT). To this end, Kruskal-Wallis tests on nominal pitch differences for each tuning system were carried out. The test was highly significant ($\chi^2=170.6$, $p<0.001$) due to the large number of pitches (N=6,381), but the median of absolute differences to nominal pitch differences between the tuning systems are actually very small (Med_ET = 15.3 cents; Med_JI = 16.8 cents; Med_PT = 15.8 cents) with comparably large standard deviations (SD_ET = 15.9 cents; SD_JI = 16.8 cents; SD_PT = 16.2 cents). Thus, the choice of the tuning system does not matter for most of the subsequent analyses, except for analysis of chord consonances (cf. below). For sake of simplicity and because it actually provides the best fit, we will use ET as a reference for the remainder of the analysis.

Finally, we checked for the global tuning of the three conditions by calculating the mean value of all C4's in the score with a nominal MIDI pitch value of 60. The mean differences to the nominal pitch are -3.7 cents, -2.15 cents, and -0.89 cents resp. for conditions AC1, AC2, and

AC3, which are sufficiently close to zero and to each other, so that global tuning corrections seemed not necessary.

B. Intonation of single pitches

We proceed in defining absolute pitch error (APE) as the absolute value of the difference of the sung pitch to the nominal value in ET (cf. Mauch et al., 2014). Pitch consistencies (PC) are defined as the standard deviation of measured pitch values for a note sung by more than one singer or sung by one singer more than once. More formally, let p_I^k be the sung pitch of a note k, where the index I = I(C, S, V) enumerates condition, singer and voice group. Let p_o^k be the nominal pitch of note k. The APE is the value $| p_I^{k-} p_o^k|$ which can be averaged across singers, notes, condition or voice group. Moreover, let $q_{S,V}{}^k$ be the average pitch for note k across singer or voice group, where the index k enumerates notes in the score for voice groups but identical notes in a voice for singers. Then the pitch consistency for note k is defined as

$$PC_{S,V} = \surd\ (1/N_{S,V}\ \Sigma s,v\ (p_I^k - q_{S,V}{}^k)^2, \tag{1}$$

i.e. the sample standard deviation of pitches in a group.

For testing the influence of acoustical feedback condition on these intonation measures, we first calculated mean APE (MAPE) as well as mean PC (MPC) per singer in each condition and subjected these values to a Friedman test with acoustical condition as block and singer as group variable. No significant differences between conditions for MAPE and MPC (all p=n.s., cf. Table I) were revealed. However, we found highly significant differences between singers (using acoustical condition as group and singer as block variable, all $p<0.001$) (see Fig. 1 for individual differences).

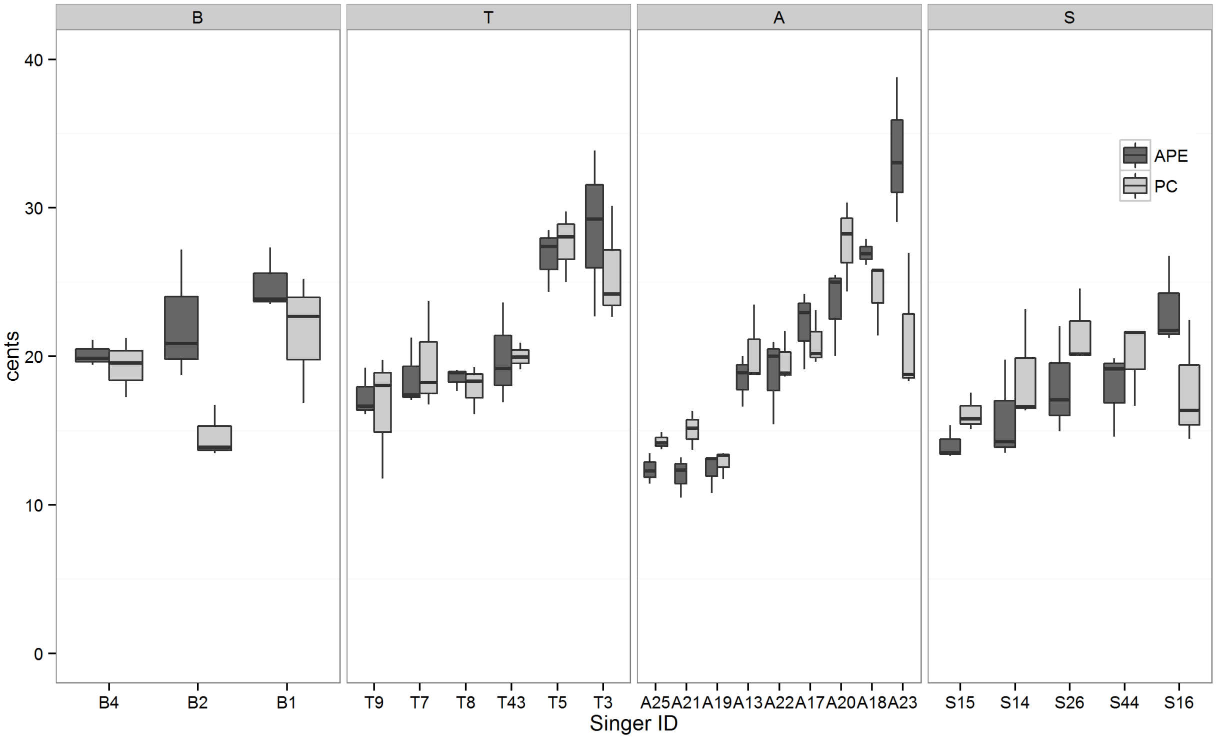

Fig. 1. Boxplots of absolute pitch errors (APE; dark grey) and pitch consistency (PC; light gray) by singer and voice group (B = Bass, T = Tenor, A = Alto, S = Soprano). APE and PC are strongly correlated (r= .718, p<0.001). APE: AM=20.3 cents, SD=5.9 cents, MIN= 10.5 cents (Alto 21, AC2), MAX=38.8 cents (Alto 23, AC3). PC values: AM=19.7 cents, SD=4.6, MIN=11.7 (Alto 19, AC3), MAX=30.3 cents (Alto 20, AC3).

The singer with the lowest APE was alto 25 with a median APE of 12.3 cents across all conditions, whereas the singer with the highest APE was alto 23 with a median APE value of 33 cents. From the partly large differences between APE and PC, it can be concluded that some singers show tendencies to sing consistently sharp or flat, but that most singers just produce random pitch errors.

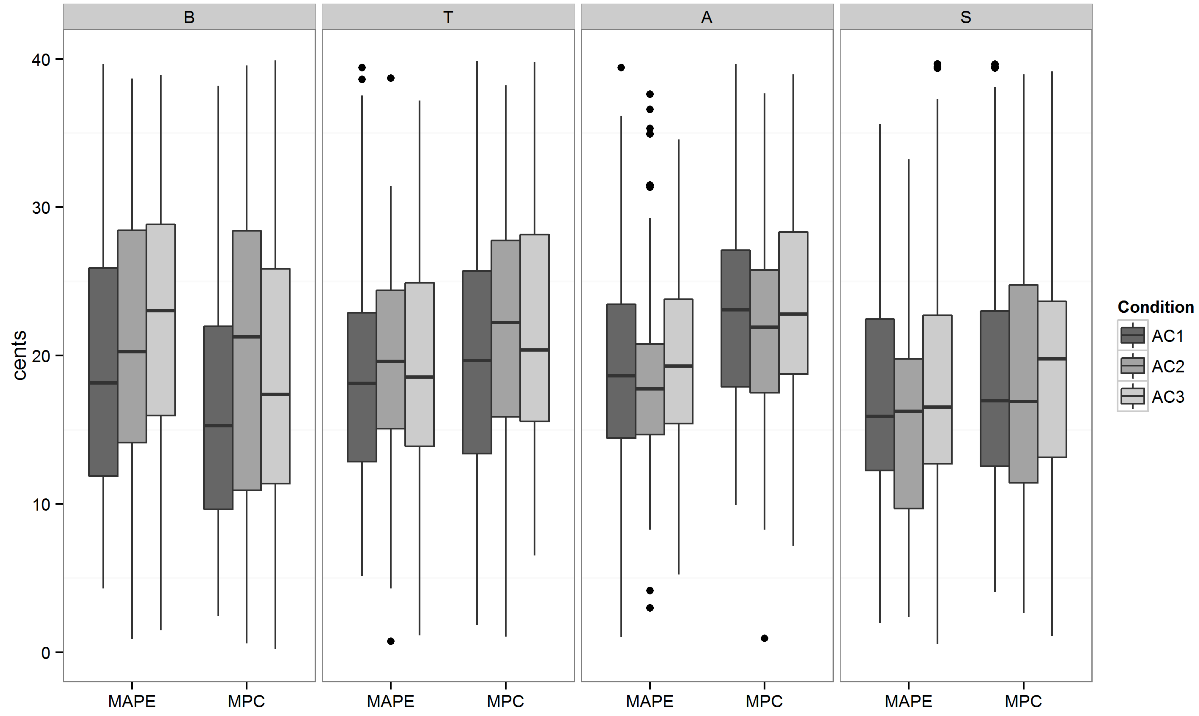

Fig. 2. Boxplots of mean absolute pitch errors (MAPE) and pitch consistency (MPC) by voice group and condition. MAPE and PC are strongly correlated (r= .66, p<0.001). MAPE values: AM=19.8 cents, SD=9.19 cents, MPC values: AM=21.4 cents, SD=9.91.

Likewise, we tested MAPE and MPC on the level of voice groups (Tab. II and III, Fig. 2) by a Friedman test over condition using note number as grouping variable. Only the value MAPE for soprano became significant (χ^2= 6.907, p=0.032), where AC3 has the largest MAPE of 18.7 cents (SD=9.1) compared to 17.6 for AC1 (SD=7.3) and 16.4 for AC2 (SD=7.9) with small effect sizes (d_{12}=-0.16, d_{13}=0.13, d_{23} = 0.28). For MPC, only the bass group became almost significance (p=0.078).

C. Intonation of consonance

We also looked at the intonation of consonances between all voice-groups of the choir. To this end, we identified about 60 synchronization points in the score where all four voices start a new tone simultaneously. Next, we calculated the mean pitch for each voice group for each synchronization point and the intervals between all voice combinations. We did the same with the nominal pitches, and calculated the mean of absolute differences of all different possible interval combination to their nominal values. This measure will be called mean interval matrix deviation (MIMD). More formally, if p_S, p_A, p_T, p_B are the pitches of soprano, alto, tenor and bass for a metrical position, then the interval matrix for this consonance at this point is defined as $d_{KL} = p_K- p_L$. Writing p^0_S, p^0_A, p^0_T, p^0_B for the corresponding nominal pitches and d^0_{KL} for the corresponding interval matrix, the MIMD is defined as

$$\text{MIMD}= 1/12\ \Sigma_{KL}|\ d_{KL}- d^0_{KL}|. \tag{2}$$

The factor 1/12 is chosen due to the fact that the absolute difference matrix is symmetric with zero diagonal. This is inessential for comparison, but facilitates interpretation of absolute values. We conducted three Friedman tests with condition as blocks (N=3) and synchronization points as groups (N=57, only points with sufficient data available could be used) for three different tuning systems. Although we argued earlier that on the level of raw pitch height the deviations from either tuning system are not discernible, we cannot rule out that singers employ locally small pitch adjustments to maximize overall consonance of a chord, as for instance proposed by Howard (2007b). The Friedman tests are not significant for the Pythagorean tuning system (χ^2= 5.660, p=0.056), (see Table IV). In Fig. 3, histograms and probability densities for the MIMD (ET) per condition are depicted. At least, it seems that chord consonances are influenced by the tuning system as well as by condition, but no clear pattern emerges. Only a broadening of MIMD values with increasing VRS can be observed, which indicates that singers might be disturbed in their micro-adjustment by the acoustical feedback, becoming less consistent in their chord intonation with increasing VRS. An Ansari-Bradley test (non-parametric F-test) revealed that for ET and PT the scale parameters in condition AC1 and AC3 are indeed significantly different (ET: p=0.034, JI: p=0.362, PT: p=0.022), as well if pooled across all tuning systems (p<0.001).

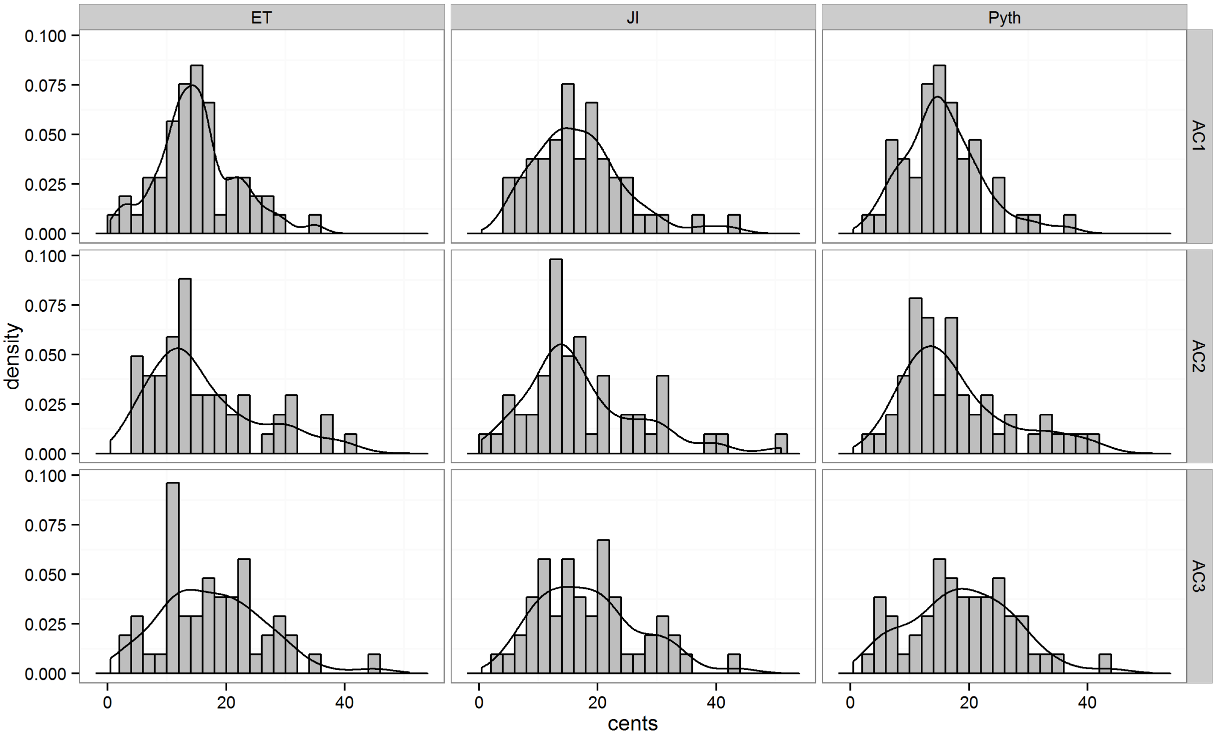

Fig. 3. Histograms and density plots of mean interval matrix deviations for different tuning systems and conditions. The spread of the distributions clearly increases with increasing virtual room size (VRS) for all tuning systems.

D. Timing and tempo

The next part of the analysis dealt with timing information for the singers in and between voices. For in-voice comparison, the fact is used, that each singer in the voice group is supposed to sing the same note. Denote for singer i in voice-group K ∈ (S, A, T, B) the onset of note number n as $t^i_K(n)$. Then define for each note in a voice group the timing precision $P_K(n)$ as the logarithm of the (sample) standard deviation of note onsets

$$P_K(n) = \log SD(t^i_K(n)). \tag{3}$$

This is defined only for notes for which at least 2 tone events per voice are annotated. The logarithm is introduced here only for reasons of better display, since standard deviations are positive values with heavily tailed distributions. The non-parametric rank tests are not affected by this strictly monotonic transformation. However, effect sizes are calculated without taking the logarithm. For timing precision across-voices, the same idea applies but only for notes at the identified synchronization points, as in the case of chord accuracy above. Hence, $P_X(s) = \log SD(t^i_K(s))$ for $s \in S$. Friedman tests were carried out to check for the influence of acoustical condition on timing precision using metrical position as grouping variable. First, across all singers (Tab. V, Fig. 4), and, second, for each voice group separately (Tab. VI and VII, Fig. 5). For raw onsets, the Friedman test became highly significant across all singers (χ^2= 31.6, p<0.001, d_{12} = -0.16, d_{13} = -0.42, d_{23} = -0.23), with decreasing timing precision for increasing VRS. However, it can be suspected that this might be mainly a tempo effect, since the tempos were quite different (mean tempo AC1: 84 bpm; AC2: 79.5 bpm; AC3: 71.7 bpm). Indeed, using normalized onsets by scaling the onsets to the interval 0–1 for each condition, the significant differences disappear (χ^2 = 2.7, p = 0.259). This is in accordance with Weber's law as well as Wing & Kristofferson's model (1973) that the SD of produced intervals should scale inversely with tempo (McAuley, 2010). But this still leaves the strong effect that the conductor chose increasingly slower tempos with increasing room size. Moreover, perceivable differences are based on absolute, not relative timing, so timing precision is in fact deteriorating, simply because it is harder to achieve synchrony in slower tempo.

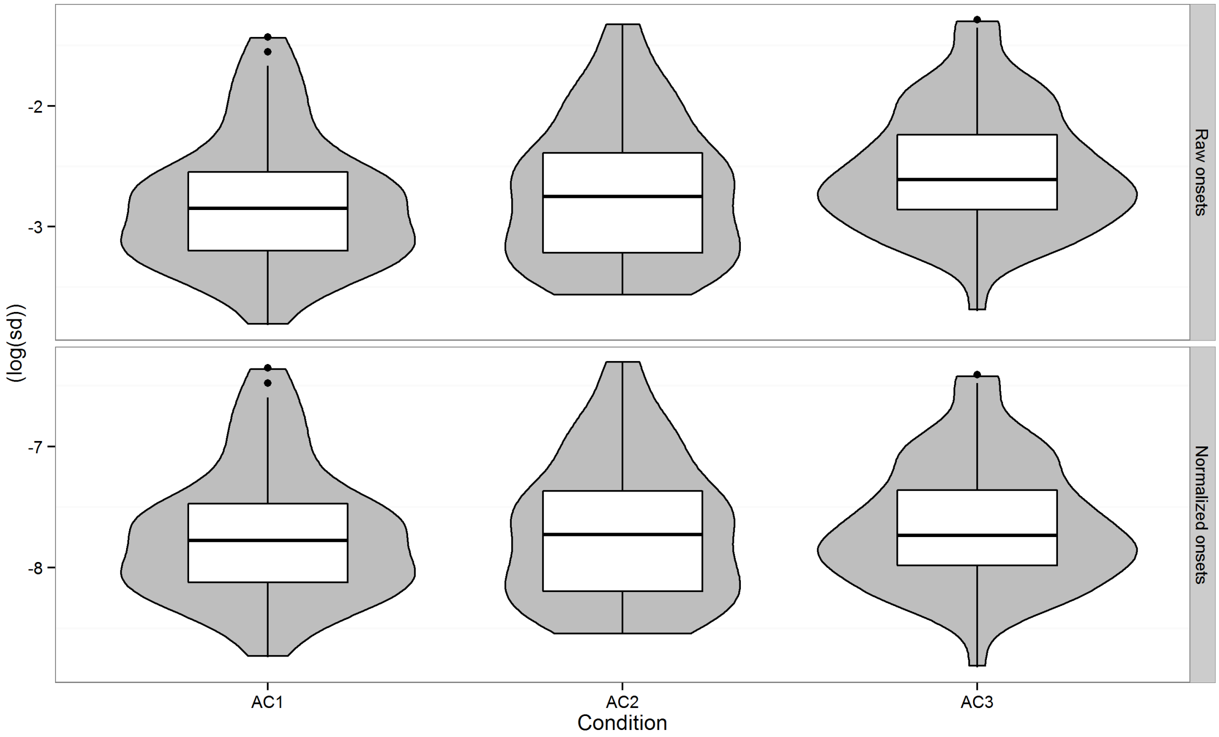

Fig. 4. Distributions of timing precision per voice group across different conditions. The upper panel shows distributions based on raw onsets; the lower panel uses normalized onsets, where the mean onset of the first note is mapped to 0 and the mean onset of the last note is mapped to 1, thus, compensating for tempo differences.

Next, we checked differences in each voice group using Friedman tests (Tab. VI and VII, Fig. 5). Using raw onsets, only the tenor and alto group were strongly influenced (tenor: χ^2 = 47.8, p<0.001, alto: χ^2 = 26.4, p<0.001), the soprano less so, but still significant (χ^2 = 8.18, p=0.017), whereas the bass seems basically unaffected (χ^2 = 5.646, p = 0.059). The largest effect size is between condition AC1 and AC3 for tenor of d_{13} = -0.61). But again, using normalized onsets (Tab. VII, Fig. 5), the significant differences for alto and soprano disappear but still persist for the tenor group. Surprisingly, the bass group showed a strong effect (χ^2 = 15.25, p<0.001), but in the opposite direction with precision actually being higher in AC3 than in the two other conditions ($d_{12} = 0.06$, $d_{13} = 0.28$, $d_{23} = 0.31$).

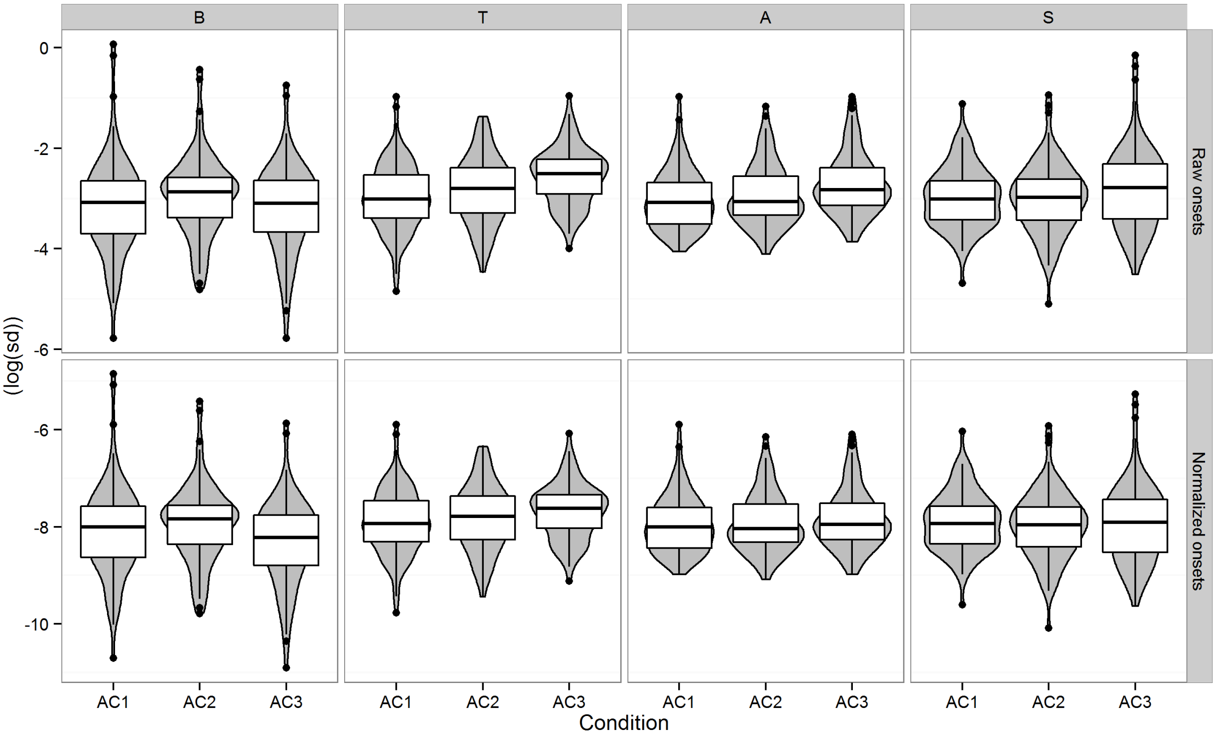

Fig. 5. Timing precision of voice groups for different feedback conditions. The upper panel shows distributions based on raw onsets, the lower panel uses normalized onsets where the mean onset of the first note is mapped to 0 and the mean onset of the last note is mapped to 1, thus, compensating for tempo differences.

E. Loudness

To assess loudness differences between the three conditions, we used the full (dry and raw) mix of the recordings instead of single tracks. We segmented the piece “Locus Iste” into 19 musical phrases, in which the full choir was singing, i.e., disregarding those parts where the tenor was leading a call-and-response section roughly in the middle of the piece. To estimate the power for each phrase, we used the smoothed power curve from the “Power Curve” plugin for Sonic

Visualiser by the Mazurka project (cf. http://www.mazurka.org.uk/software/sv/plugin/). Results are shown in Fig. 5. One clearly sees, that nearly consistent over all phrases, AC2 was sung the loudest (median power -23.1 dB), followed by AC1 (median power -24.5 dB) and AC3 (median power -25.0 dB). A Friedman test of median power per phrase across conditions with phrases as grouping variable showed that these differences are highly significant ($\chi^2 = 17.16$, $p<0.001$), but that the effect sizes are rather small ($d_{12} = -0.20$, $d_{13} = 0.07$, $d_{23} = 0.27$).

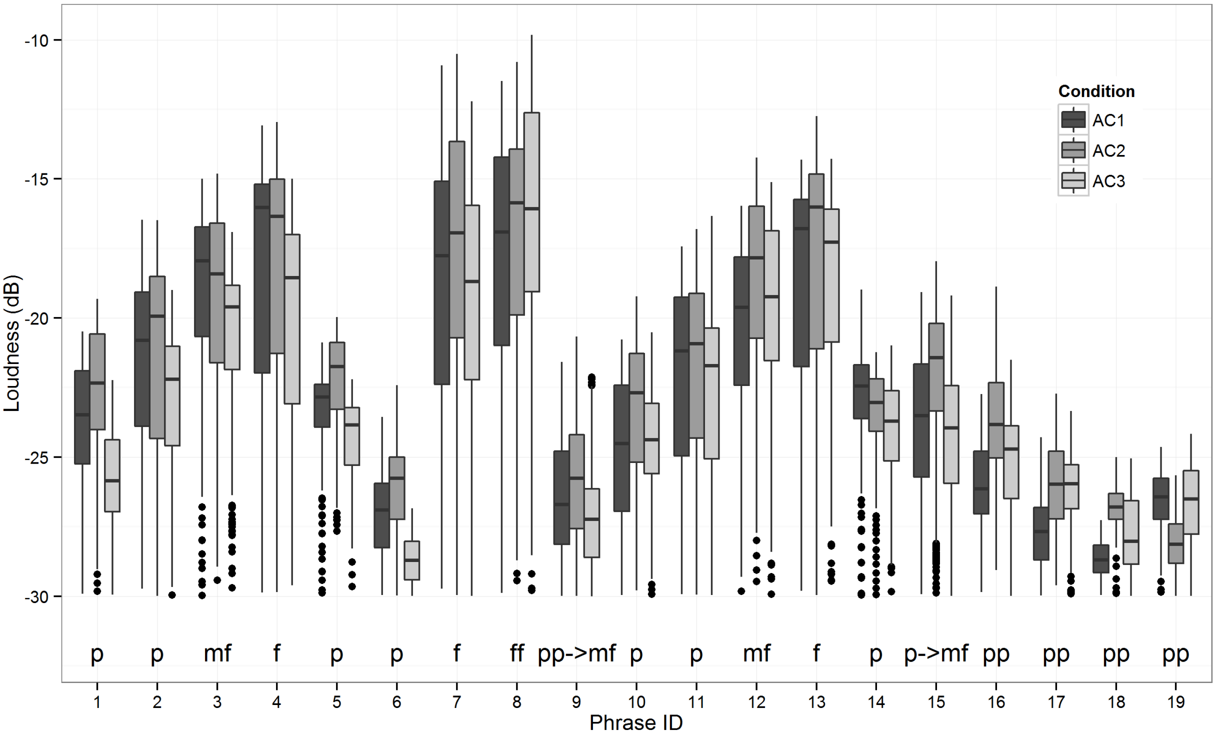

Fig. 6. Intensity values per phrase across different conditions. AC2 was sung with the loudest, AC3 with the lowest volume. Loudness values were cut off at -30 dB for sake of display.

F. Subjective measures

In Fig. 7 the results of the short questionnaire are depicted. For nearly all items, condition AC2 showed clearly the largest preference. Only for the item “Easy to sing in time”, the values decreased with increasing VRS. Moreover, the singers were rather discouraged to sing in the dry condition AC1 (Item “Encouraged to Sing”, AM = 2.74, SD = 1.25). For the most difficult condition AC3 encouragement was rated as being much higher (AM = 3.34, SD = 1.27). For all other items, the mean values for condition AC1 and AC3 were about the same magnitude and each smaller than AC2. Differences for the six subjective variables were mostly significant or highly significant with respect to condition according to multiple Kruskal-Wallis tests (Table VIII), even after Bonferroni correction for multiple testing. Effect sizes were sometimes very large, e.g., for “Easy to sing” and “Encouraged to sing” in AC1 vs. AC2 with values d_{12} = 1.393 and 1.344 respectively. The largest effect size with a value of d_{13} = -2. 373 was found between condition AC1 and AC3 for the variable “Easy to sing in time”. Introducing a rather moderately VRS in condition AC2 resulted already in a decline in experienced rhythm precision with d_{12} = -0.6, but the difference between AC2 and AC3 is even much more dramatic with d_{23} = -1.949. The variable concerning the SOR, “Easy to hear to oneself” and “Easy to hear others” are much less affected by VRS, the former showing no significant difference even before Bonferroni correction (p=.228). Likewise, the item “Easy to sing in tune” was barely affected with the largest effect between condition AC2 and AC3 of d_{23} = -0.959.

To check for connections between the subjective assessments and objective measures, we performed Spearman’s rank correlations of the six subject items with MAPE, MPC, mean onset differences (MDD), absolute mean onset differences (AMDO) and standard deviation of onset difference (SDO), where onset differences were calculated with respect to the mean onset of each tone in a voice group. Only a few correlations became significant. Across all conditions, MPC correlated negatively with the variable “Easy to hear oneself” (ρ= -0.248, p = 0.039), hence the better the impression to hear oneself, the better the pitch consistency. Similarly, MPC correlated negatively with “Easy to sing in time” (ρ = -0.244, p = 0.044). SDO correlated negatively with “Easy to hear oneself” (ρ= -0.242, p = 0.045) and with “Easy to sing in time” (ρ = -0.307, p =

0.01), i.e. singers estimating higher difficulties with timing, were in fact less consistent in their timing.

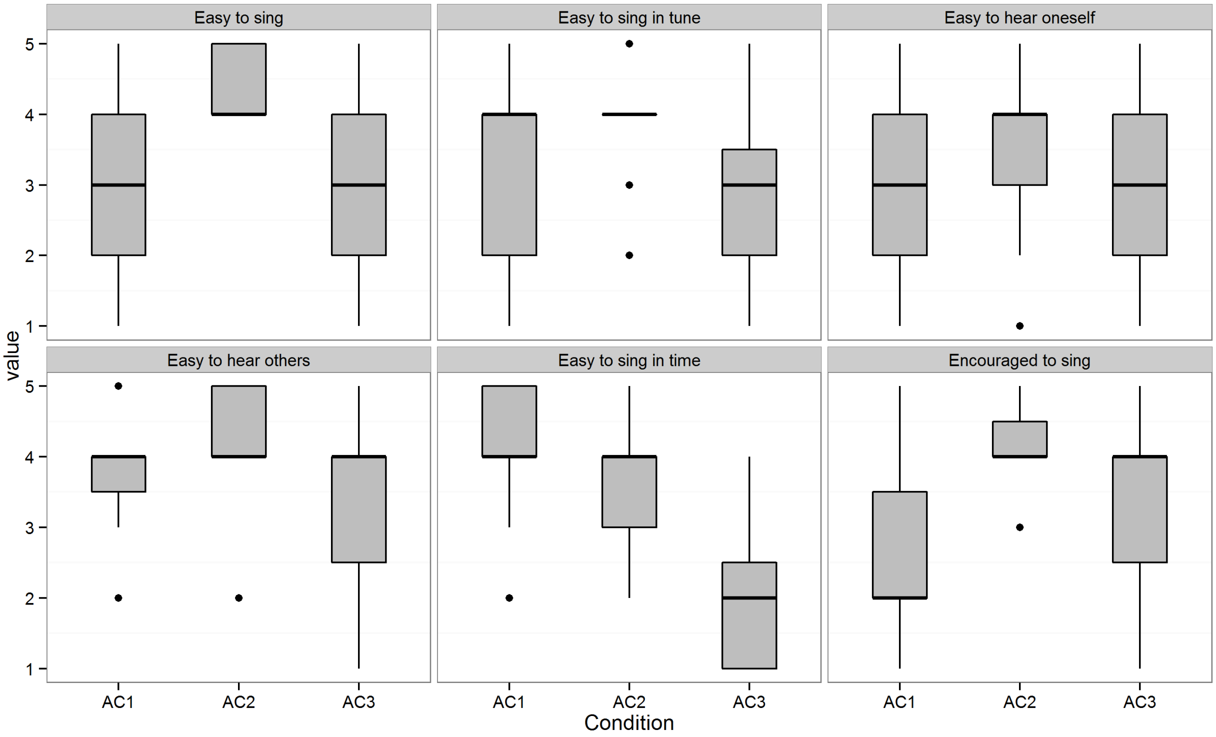

Fig. 7. Boxplots of subjective evaluations of performances in different conditions. All scales 5-point Likert scales from 1=totally disagree to 5=totally agree. AC2 was the most comfortable condition, except for the variable "Easy to sing in time" whereas AC1 was the most preferred.

The puzzling correlation of the timing related item "Easy to sing in time" with pitch consistency may be explainable by some other interesting correlations: MPC with SDO (ρ= 0.312, p = 0.008) and MPC with MDO (ρ = 0.249, p = 0.038). Moreover, MAPE correlated with MDO, AMDO and SDO as well. It seems that the best singers with regard to intonation also take the lead, i.e., singing ahead of the rest of the voice group and show also a more consistent timing. Actually the correlation of MAPE and MDO was the strongest overall (ρ= 0.366, p = 0.002), disregarding the strong correlation of MAPE and MDO. Finally, all questionnaire items were strongly correlated with each other (mean Pearson correlation r = 0.59) and singers were rather consistent in their ratings (Cronbach's alpha over all items and condition α = 0.84).

IV. DISCUSSION

The main outcome of the present study with respect to objective measures is that tempo tends to be notably slower and timing is less precise for increasing virtual room size (VRS), whereas intonation is only weakly influenced. On the other hand, our results revealed that subjective experience was much more affected with a clear preference for the medium VRS of AC2. The largest VRS of AC3 resulted in a mean encouragement drop below the neutral value, i.e., indication of a de-motivational effect. All objective effects are rather small, and might be even practically irrelevant (few cents, few millisecond, though this has to be tested in future perception experiments). At least, they seem to not be as relevant as the decrease in subjective singing comfort with very dry and very wet conditions. Even though it is still not clear why, in general, humans seem to prefer sounds with a small to medium amount of reverberation, a possible explanation for our special case might be given. First, the lesser preference for the dry condition might be due to the "unnatural" impression of sounds without any acoustical room added. Second, the even weaker preference for the very large (virtual) room (AC3) might be due to the increase in singing effort required to compensate for weakened rhythmical precision, which in turn might result from blurred onsets or the slower tempo. In turn, the slower tempo might have been chosen by the conductor intuitively to keep word intelligibility constant (Harris & Reitz, 1985) as well as to avoid undesired dissonances by fusion of direct sound and reflections.

The finding that the choir tends to sing slower with increasing VRS is in accordance with results reported by Schärer Kalkandjiev and Weinzierl (2015) and Ueno et al. (2010). It is also in agreement with the recommendations by Quantz (1752). He emphasized the beneficial effect of playing slower in large rooms compared to playing in small rooms in order to preserve the intelligibility of the music.

The significant differences (with small to medium effect sizes) between the three acoustical conditions with regard to timing precision across all singers (for raw onsets) may be, indeed, based on a tempo effect, since differences largely, but not fully disappear using normalized onsets in accordance with Weber's law (i.e., standard deviation proportional to tempo).

Intonation analyses showed that, across the board, effect sizes were nearly all small. Hence, the objectively measurable influence of reverberation is rather on the subtle side. No significant differences between the three different conditions for mean absolute pitch error (MAPE) and mean pitch consistency (MPC) could be found (despite being highly different across singers) when looking at single pitches by singer. Likewise, no general trend could be found for consonances, even though the mean interval matrix deviation tend to be less stable (i.e., showing larger variances) with increasing VRS.

Finally, it could be observed that singers consistently sung louder in condition AC2 compared to AC1 and AC3. This might be caused either by cognitive or emotional effects or a combination thereof. On the cognitive side, the singer’s own voice may sound louder to him or her than the reference sound from the choir in AC1. In contrast, the intensity level in AC3 may be perceived as louder because of the very long RT (cf., Ternström, 1989), which might result in a tendency of the singer to sing more softly, combined with, or elicited by, a feeling of uncertainty. On the emotional side, when the artificial acoustics changed to ”very good” (AC2) the feeling of openness of space, good sound quality of the choir, and ease of singing increased, resulting in the louder singing and overall heightened expressiveness of the choir. This is also reflected in the value of the subjective item “Encouraged to sing”. This emotional effect of an “optimal” room acoustics is in line with findings from research in artificial virtual environments (Tajadura-Jimenez, Larsson, Valjamae, Vastfjall, & Kleiner, 2010; Västfjäll & Larsson, 2002), where a medium sized virtual acoustical room was shown to enhance pleasantness and arousal judgments of neutral musical and other sound stimuli. However, the reasons for this preference and emotional effect still remain unclear.

Overall, the analyses of the intonation data suggest an optimal choir singing room size of condition AC2, at least with regard to this particular piece of music ("Locus Iste") and this particular choir, though we have no doubt on the generalizability of this result. Thereby, our findings support the idea that choir conductors should always be aware of room acoustics. For instance, regarding which room would provide the best acoustical environment for a given musical piece. With respect to motivational effects, choirs might strive to perform and to practice in rooms with preferable acoustical conditions.

In addition to acoustical differences between artificial and real acoustics, it is important to keep in mind that many other parameters, not under consideration in this study, may also have an impact on the quality of singing performance. These include environmental/visual cues, architecture of the venue, reactions of the audience, social relationships/interactions between choir singers, and the actual performance situation (concert, matinee, etc.). Last but not least, the influence of different VRSs also depends on the features of music performed. For example, a fast piece with a lot of short notes is probably much more affected by large RTs and reflections, than a slow piece with many long notes.

V. Outlook

Although choir singing is one of the most frequent musical activities in the world, research on the acoustics of choir singing is quite rare. This lack of research might be due to the complexity of the examination object in itself as well as the inherent demands when analyzing multiple voice recordings (Ternström et al., 2012).

Most studies within this research area typically focus on the investigation of single voices by recording and analyzing choir singers individually in order to keep control over the dependent variables. Performance analyses of choirs based on individual recordings of each singer remain a fairly sophisticated challenge, since there is a large amount of influencing factors.

However, the choir study presented here may be exemplary for future research on choir acoustics in search of verified knowledge about the complex interactions between voices of a choir and room acoustical influences during actual performance. In order to reach this goal, further research on choir singing should be conducted using multitrack recordings under different acoustical conditions. These include acoustical environments simulated by dynamic binaural synthesis in an anechoic room using extra-aural headphones (cf., Schärer Kalkandjiev and Weinzierl (2015) or high-fidelity sound field simulation of different virtual acoustical situations similar to controlled lab experiments but without using headphones. Conversely, it would be desirable to replicate our experiment with choral performances under realistic conditions in different music venues (Bonsi et al., 2013). It would also be of interest to combine objective performance analyses with a subjective reception task on the aesthetic appreciation of choir performances by asking experts as well as non-experts to evaluate different versions of systematically manipulated multitrack recordings.

ACKNOWLEDGMENTS

We wish to thank the choir singers of the choir “Studiokuoro” from Jyväskylä (Finland) and the conductor Hannu Ikonen from the JAMK University of Applied Sciences (Jyväskylä) for their participation; also Mikko Leimu, recording engineer from the Studio Musica at the

Department of Music, University of Jyväskylä. Special thanks are given to Olivier Lartillot and Martin Hartmann who helped in data analysis in the first steps of the paper. This research was supported by the University of Jyväskylä (Finland).

[1] The Pre Delay is defined as the time between the direct sound and the first reflection(s). Increasing the Pre Delay usually changes the perceived clarity of the sound or vocals. RTs are defined as the duration to drop 60 dB below the original level (Ternström & Karna, 2002).

[2] Several other methods for automatically extracting performance data or computer-aided melody note transcription tools for the analysis of single track voice recordings have been proposed (Devaney, Mandel, Ellis, & Fujinaga, 2011).

Bonsi, D., Boren, B., Howard, D., Longair, M., Moretti, L., & Orlowski, R. (2013). Acoustic and audience response analyses of eleven Venetian churches. *Acoustics in Practice, 1*(1), 39-52.

Coleman, R. F. (1994). Dynamic intensity variations of individual choral singers. *Journal of Voice, 8*(3), 196-201.

Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2007). Singing proficiency in the general population. *J Acoust Soc Am, 121*(2), 1182. doi: 10.1121/1.2427111

Daugherty, J. F. (2003). Choir spacing and formation: Choral sound preferences in random, synergistic, and gender-specific chamber choir placements. *International Journal of Research in Choral Singing, 1*(1), 48-59.

Declercq, N. F., & Dekeyser, C. S. A. (2007). Acoustic diffraction effects at the Hellenistic amphitheater of Epidaurus: Seat rows responsible for the marvelous acoustics. *J Acoust Soc Am, 121*(4), 2011. doi: 10.1121/1.2709842

Devaney, J., Mandel, M. I., Ellis, D. P. W., & Fujinaga, I. (2011). Automatically extracting performance data from recordings of trained singers. *Psychomusicology: Music, Mind and Brain, 21*(1-2), 108-136. doi: 10.1037/h0094008

Farnetani, A., Prodi, N., & Pompoli, R. (2008). On the acoustics of ancient Greek and Roman theatres. *J Acoust Soc Am, 124*(3), 1557-1567.

Fischinger, T., & Hemming, J. (2011). Wie singe ich im Chor? - Intonations- und Timingmessungen in Vokalensembles [How do I sing in the choir? - Intonation and Timing Measurements in Vocal Ensembles]. In T. Greuel, U. Kranefeld, & E. Szczepaniak (Eds.), *Singen und Lernen* [*Singing and Learning*]. (pp. 45-58): Aachen: Shaker.

Grell, A., Sundberg, J., Ternström, S., Ptok, M., & Altenmüller, E. (2009). Rapid pitch correction in choir singers. *J Acoust Soc Am, 126*(1), 407-413. doi: 10.1121/1.3147508

Harris, R. W., & Reitz, M. L. (1985). Effects of Room Reverberation and Noise on Speech Discrimination by the Elderly. *Audiology, 24*, 319-324.

Howard, D. M. (2007a). Equal or non-equal temperament in a capella SATB singing. *Logoped Phoniatr Vocol, 32*(2), 87-94. doi: 10.1080/14015430600865607

Howard, D. M. (2007b). Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation. *J Voice, 21*(3), 300-315. doi: 10.1016/j.jvoice.2005.12.005

Jers, H., & Ternström, S. (2005). Intonation analysis of a multi-channel choir recording. *TMHQPSR Speech, Music and Hearing: Quarterly Progress and Status Report, 47*(1), 1-6.

Kircher, A. (1650). *Musurgia universalis [Music Encyclopedia].* Rome: Francesco Corbelletti (Vol. I) and Ludovico Grignani (Vol. II), book IX, Part IV.

Marshall, A. H., & Meyer, J. (1985). The directivity and auditory impressions of singers. *Acustica, 58*, 130-140.

Mauch, M., & Dixon, S. (2014). pYIN: a fundamental frequency estimator using probabilistic threshold distributions. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 659–663).

Mauch, M., Frieler, K., & Dixon, S. (2014). Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory. *J Acoust Soc Am, 136*(1), 401-411.

McAuley, J. D. (2010). Tempo and Rhythm. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music Perception* (Vol. 36, pp. 165-199). New York: Springer.

Molina, E., Tardón, L. J., Barbancho, I., & Barbancho, A. M. (2014). The importance of F0 tracking in query-by-singing-humming. *Proceedings of the 15th International Society for Music Information Retrieval Conference* (pp. 277–282).

Mozart, L. (1756). *Versuch einer gründlichen Violinschule [A Treatise on the Fundamental Principles of Violin Playing]*. Augsburg: Johann Jakob Lotter, §.8., p. 221.

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *J Acoust Soc Am, 128*(4), 2182-2190. doi: 10.1121/1.3478782

Pörschmann, C. (2000). Influences of bone doncution and air conduction on one's own voice. *Acustica/Acta Acustica, 86*(6), 1038-1045.

Quantz, J. J. (1752). *Versuch einer Anweisung, die Flöte traversière zu spielen [Attempt of an instruction to play the flute traversière]*. Berlin: Johann Friedrich Voß, Chap. XVI, §18, p. 170.

Schärer Kalkandjiev, Z., & Weinzierl, S. (2013). The Influence of Room Acoustics on Solo Music Performance: An Empirical Case Study. *Acta Acustica united with Acustica, 99*(3), 433-441. doi: 10.3813/aaa.918624

Schärer Kalkandjiev, Z., & Weinzierl, S. (2015). The Influence of Room Acoustics on Solo Music Performance: An Experimental Study. *Psychomusicology: Music, Mind, and Brain*. doi: 10.1037/pmu0000065

Seaton, R., Sharp, D., & Pim, D. (2014). Pitch drift in a capella choral singing: the outcomes from an international survey. In *Institute of Acoustics 40th Anniversary Conference*, Birmingham, UK, (pp. 312–319).

Sundberg, J. (1987). *The science of the singing voice*. Dekalb, IL: Northern Illinois University Press, Chap. 6.

Tajadura-Jimenez, A., Larsson, P., Valjamae, A., Vastfjall, D., & Kleiner, M. (2010). When room size matters: acoustic influences on emotional responses to sounds. *Emotion, 10*(3), 416-422. doi: 10.1037/a0018423

Ternström, S. (1989). Long-time average spectrum characteristics of different choirs in different rooms. *STL-QPSR, 3*, 15-31.

Ternström, S. (1993). Long-time average spectrum characteristics of different choirs in different rooms. *Voice (UK), 2*, 55-77.

Ternström, S. (1994). Hearing myself with others: Sound levels in choral performance measured with separation of one's own voice from the rest of the choir. *Journal of Voice, 8*(4), 293-302.

Ternström, S. (1999). Preferred self-to-other ratios in choir singing. *J Acoust Soc Am, 105*(6), 3563-3574.

Ternström, S. (2003). Choir acoustics–an overview of scientific research published to date. *International Journal of Research in Choral Singing, 1*(1), 3-12.

Ternström, S., Cabrera, D., & Davis, P. (2005). Self-to-other ratios measured in an opera chorus in performance. *J Acoust Soc Am, 118*(6), 3903. doi: 10.1121/1.2109212

Ternström, S., Jers, H., & Nix, J. (2012). Group and Ensemble Vocal Music. In G. McPherson & G. F. Welch (Eds.), *The Oxford Handbook of Music Education,* (Vol. 1, pp. 580-593). Oxford: Oxford University Press.

Ternström, S., & Karna, D. R. (2002). Choir. In R. Parncutt & G. McPherson (Eds.), *The Science and Psychology of Music Performance*, (pp. 269-283). New York: Oxford University Press.

Ternström, S., & Sundberg, J. (1988). Intonation precision of choir singers. *J Acoust Soc Am, 84*(1), 59-69.

Ueno, K., Kato, K., & Kawai, K. (2010). Effect of Room Acoustics on Musicians' Performance. Part I: Experimental Investigation with a Conceptual Model. *Acta Acustica united with Acustica, 96*(3), 505-515. doi: 10.3813/aaa.918303

Västfjäll, D., & Larsson, P. (2002). Emotion and Auditory Virtual Environments: Affect-Based Judgments of Music Reproduced with Virtual Reverberation Times. *Cyber Psychology & Behavior, 5*(1), 19-32.

Vitruvius, M. (1st Century BC). *De Architectura [On architecture],* book V, Public places, Chap. 3.

Wing, A. M., & Kristofferson, A. B. (1973). The timing of interresponse intervals. *Perception & Psychophysics, 13*(3), 455-460.

Zarlino, G. (1558). *Le istitutioni harmoniche [Treatise on music theory and performance practice]*, Venice: F. de Franceschi, Part III, Chap. 66.

Tables

TABLE I. Descriptive statistics and Friedman tests for mean absolute pitch error (MAPE) and mean pitch consistency (MPC) per condition evaluated for individual singers. All means and standard deviations given in cents; all degrees of freedom=2.

	MAPE		**MPC**	
	AM	**SD**	**AM**	**SD**
AC1	20.4	6.1	19.3	4.9
AC2	19.6	5.6	19.2	4.3
AC3	20.8	6.1	20.7	4.6
χ^2	4.174		1.130	
p	0.124		0.568	

TABLE II. Descriptive statistics and Friedman tests for mean absolute pitch error (MAPE) per note for voice groups. All means and standard deviations given in cents; all degrees of freedom = 2.

	Bass		**Tenor**		**Alto**		**Soprano**	
	AM	**SD**	**AM**	**SD**	**AM**	**SD**	**AM**	**SD**
AC1	21.3	10.8	19.6	8.8	19.5	7.1	17.6	7.3
AC2	22.9	11.4	20.1	7.7	18.3	5.9	16.4	7.9
AC3	23.2	10.0	19.9	8.7	20.4	7.2	18.7	9.1
N	87		107		99		108	
χ^2	3.057		1.701		4.384		6.907	
p	0.217		0.427		0.112		0.032*	

TABLE III. Descriptive statistics and Friedman tests for mean pitch consistency (MPC) per note evaluated for voice groups. All means and standard deviations given in cents; all degrees of freedom=2.

	Bass		**Tenor**		**Alto**		**Soprano**	
	AM	**SD**	**AM**	**SD**	**AM**	**SD**	**AM**	**SD**
AC1	19.2	12.7	21.5	10.2	23.9	7.5	18.6	8.6
AC2	21.6	11.9	22.6	8.6	21.7	6.2	18.8	9
AC3	20.9	11.8	23.4	10.2	23.3	6.7	21	9.9
N	87		107		99		108	
χ^2	5.186		4.392		3.735		2.29	
p	0.075		0.111		0.155		0.318	

TABLE IV. Descriptive statistics and Friedman tests for chord consonances with respect to tuning and condition using metrical position as grouping variable. All means and standard deviations given in cents; all degrees of freedom=2.

	ET		**JI**		**Pyth**	
	AM	**SD**	**AM**	**SD**	**AM**	**SD**
AC1	14.8	6.4	16.2	7.1	15.6	6.3
AC2	15.8	8.8	16.9	9.7	17.1	8.2
AC3	17.4	8.9	18.2	8.2	18.6	8.9
χ^2	1.149		0.894		5.660	
p	0.563		0.640		0,059	

TABLE V. Descriptive statistics and Friedman tests for timing precision across voices with respect to condition and raw/normalized onsets using metrical position as group variable. All degrees of freedom=2.

Condition	**Raw Onsets**		**Normalized Onsets**	
	AM (ms)	**SD (ms)**	**AM (x1000)**	**SD (x1000)**
AC1	70.9	45.9	0.515	0.333
AC2	79.2	52.8	0.545	0.364
AC3	91.6	53.4	0.546	0.318
χ^2	31.6		2.7	
P	0.000***		0.259	

TABLE VI. Descriptive statistics and Friedman tests for timing precision per voice group over raw onsets using with metrical position as grouping variable. N indicates number of usable points per voice group Effect sizes are estimated as differences of mean of second condition to first condition divided by mean of standard deviations.

	Bass		**Tenor**		**Alto**		**Soprano**	
	AM	**SD**	**AM**	**SD**	**AM**	**SD**	**AM**	**SD**
AC1	78.6	153.4	59.2	46.7	58.8	49.7	61.3	44.2
AC2	75.5	95.4	73.8	50.9	62.9	48.4	64.9	56.7
AC3	59.7	69.6	90.5	56.5	82.5	65.6	89.8	119.8
N	79		101		97		102	
χ^2	5.646		47.782		26.412		8.176	
p	0.059		0.000***		0.000***		0.017*	
d_{12}	0.025		-0.299		-0.084		-0.071	
d_{23}	0.192		-0.311		-0.344		-0.281	
d_{13}	0.170		-0.607		-0.412		-0.347	

TABLE VII. Descriptive statistics and Friedman tests for timing precision per voice group over normalized onsets using metrical position as grouping variable. N indicates number of usable notes per voice group Effect sizes are estimated as differences of mean of second condition to first condition divided by mean of standard deviations.

	Bass		**Tenor**		**Alto**		**Soprano**	
AC1	0.571	1.113	0.430	0.339	0.427	0.361	0.445	0.321
AC2	0.520	0.657	0.508	0.351	0.433	0.333	0.447	0.390
AC3	0.356	0.415	0.539	0.337	0.492	0.391	0.535	0.714
N	79		101		97		102	
χ^2	15.215		18.554		2.619		2.608	
p	0.000***		0.000***		0.270		0.271	
d_{12}	0.058		-0.227		-0.018		-0.005	
d_{23}	0.306		-0.092		-0.163		-0.160	
d_{13}	0.281		-0.325		-0.173		-0.173	

TABLE VIII. Kruskal-Wallis tests for the 6 subjective variables by acoustical conditions P-values significances are Bonferroni corrected. Effect sizes are given in the format $d_{Cond1Cond2.}$ All degrees of freedom df=2.

Item	**χ^2**	**p**	**d12**	**d13**	**d23**
Easy to sing	19.4	0.000***	-1.393	0.069	1.523
Easy to sing in tune	8.0	0.019	-0.333	0.452	0.959
Easy to hear oneself	2.6	0.278	-0.264	0.211	0.500
Easy to hear others	6.4	0.041	-0.578	0.283	0.795
Easy to sing in time	35.5	0.000***	0.653	2.373	1.949
Encouraged to sing	13.5	0.001**	-1.344	-0.517	0.686

Figure captions

Fig. 1. Boxplots of mean absolute pitch errors (MAPE; dark grey) and pitch consistency (PC; light gray) by singer and voice group. MAPE and PC are strongly correlated (r= .718, p<0.000). MAPE: AM=20.3 cents, SD=5.9 cents, MIN= 10.5 cents (Alto 21, AC2), MAX=38.8 cents (Alto 23, AC3). PC values: AM=19.7 cents, SD=4.6, MIN=11.7 (Alto 19, AC3), MAX=30.3 cents (Alto 20, AC3).

Fig. 2. Boxplots of mean absolute pitch errors (MAPE) and pitch consistency (MPC) by voice group and condition. MAPE and PC are strongly correlated (r= .66, p<0.000). MAPE values: AM=19.8 cents, SD=9.19 cents, MPC values: AM=21.4 cents, SD=9.91.

Fig. 3. Histograms and density plot of mean interval matrix deviation for different tuning systems and conditions. The spread of the distributions clearly increases with increasing reverberation for all tuning systems.

Fig. 4. Boxplots of mean absolute pitch errors (MAPE) and pitch consistency (MPC) by voice group and condition. MAPE and PC are strongly correlated (r= .66, p<0.001). MAPE values: AM=19.8 cents, SD=9.19 cents; MPC values: AM=21.4 cents, SD=9.91.

Fig. 5. Timing precision of voice group for different feedback conditions. The upper panel shows distribution based on raw onsets, the lower panel uses normalized onsets where the mean onset of the first note is mapped to 0 and the mean onset of the last note is mapped to 1, thus, compensating for tempo differences.

Fig. 6. Intensity values per phrased across different conditions. AC2 was sung with the loudest, AC3 with the lowest volume. Loudness value were cut off at -30 dB for sake of display.

Fig. 7. Boxplots of subjective evaluations of performances in different conditions. All scales 5-point Likert scale with 1=totally disagree to 5=totally agree. AC2 was the most comfortable condition, except for the variable "Easy to sing in time" where AC1 most the most preferred.