ANALYSIS OF VOCAL IMITATIONS OF PITCH TRAJECTORIES

Jiajie Dai, Simon Dixon
Centre for Digital Music, Queen Mary University of London, United Kingdom
{j.dai, s.e.dixon}@qmul.ac.uk

ABSTRACT

In this paper, we analyse the pitch trajectories of vocal imitations by non-poor singers. A group of 43 selected singers was asked to vocally imitate a set of stimuli. Five stimulus types were used: a constant pitch (*stable*), a constant pitch preceded by a pitch glide (*head*), a constant pitch followed by a pitch glide (*tail*), a pitch *ramp* and a pitch with *vibrato*; with parameters for main pitch, transient length and pitch difference. Two conditions were tested: singing simultaneously with the stimulus, and singing alternately, between repetitions of the stimulus. After automatic pitch-tracking and manual checking of the data, we calculated intonation accuracy and precision, and modelled the note trajectories according to the stimulus types. We modelled pitch error with a linear mixed-effects model, and tested factors for significant effects using one-way analysis of variance. The results indicate: (1) Significant factors include stimulus type, main pitch, repetition, condition and musical training background, while order of stimuli, gender and age do not have any significant effect. (2) The *ramp*, *vibrato* and *tail* stimuli have significantly greater absolute pitch errors than the *stable* and *head* stimuli. (3) Pitch error shows a small but significant linear trend with pitch difference. (4) Notes with shorter transient duration are more accurate.

1. INTRODUCTION

Studying the vocal imitations of pitch trajectories is extremely important because most of the human produce a musical tone by imitation rather than absolute. Only .01% of the general population can produce a musical tone without the use of an external reference pitch [22]. Although sing in tone is the primary element of singing performance, the research of vocal imitations with unstable stimuli has not been explored. It is significant to distinguish the influence factors and to quantise them, fill the gap between response and stimuli, as well as create knowledge to help the future music education and entertainment.

The accuracy of pitch in playing or singing is called intonation [8, 20]. Singing in tune is extremely important for solo singers and choirs because they must be accurate and blend well with accompaniments and other vocal parts [1]. However, it is a practical challenge when the singers have to sing with an unstable reference pitch or other vocal parts without instrumental accompaniment [17, Ch. 12, p. 151]. Nevertheless, most singers rely on their sense of relative pitch and their teammates who provide reference pitches which help them maintain tuning, as the initial tonal reference can be forgotten over time [9, 11]. Pfordresher *et al.* [16] distinguish between pitch accuracy, the average difference between the sung pitch and target pitch, and pitch precision, the standard error of sung pitches.

As for vocal reference pitch (stimulus of imitation in this paper), it usually does not have a fixed pitch for each note which is different from percussion instruments with a stable shape [4, 7, 11]. Instead, vocal notes typically fluctuate around the target pitch. When singing with a stable reference pitch, the singer will voluntarily adjust their vocal output until the auditory feedback matches the intended note [28]. This adjustment especially at the beginning of the note, they may sing with vibrato, and they may not sustain the pitch at the end of the note [27]. Although singers make fewer errors when singing in unison or with stable accompaniment [24], the response of unstable stimulus or notes with transient parts is still obscure.

A transient is part of a signal (often at the beginning) during which its properties are rapidly changing and thus unpredictable. For most musical tones, a short transient segment is followed by a much longer steady state segment, but for singing, such a segmentation is difficult, as the signal never reaches a steady state. At the beginning of a tone, a pitch glide is often observed as the singer adjusts the vocal cords from their previous state (the previous pitch or a relaxed state). Then the pitch is adjusted as the singer uses perceptual feedback to correct for any error in the pitch. Possibly at the same time, vibrato may be applied, which is an oscillation around the central pitch, which is close to sinusoidal for skilled singers, but asymmetric for unskilled singers [7]. At the end of the tone, the pitch often moves in the direction of the following note, or downward (toward a relaxed vocal cord state) if there is no immediately following note.

To investigate the response of singers to time-varying pitch trajectories, we prepared a controlled experiment using synthetic stimuli, in order to test the following hypotheses:

- The stimulus type will have a significant effect on intonation accuracy.
- A greater duration or extent of deviation from the

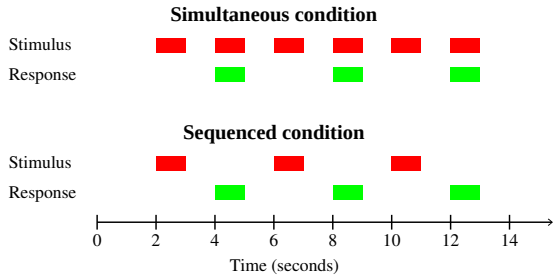

Figure 1: Experimental design showing timing of stimuli and responses for the two conditions.

main pitch will increase intonation error.

- The direction of any deviation in the stimulus from the main pitch determines the direction of any error in the response.
- Singing simultaneously with the stimulus will result in a lower error than alternating the response with the stimulus.

We extract the fundamental frequency (f_0) [5, 10] and convert to a logarithmic scale, corresponding to non-integer numbers of equal-tempered semitones from the reference pitch (A4, 440Hz). We model responses according to stimulus types in order to compute the parameters of observed responses. The significance of factors (stimulus type, stimulus parameters and order of stimuli, as well as participants' musical background, gender and age) was evaluated by analysis of variance (ANOVA) and linear mixed-effects models.

2. MATERIALS AND METHODS

2.1 Experimental Design

The experiment consisted of 75 trials in each of two conditions. In each trial, the participant imitated the stimulus three times (see Figure 1). Each stimulus was one second in duration. In the *simultaneous* condition, the stimulus was repeated six times, with one second of silence between the repetitions, and the participants sang simultaneously with the 2nd, 4th and 6th instances of the stimulus. The *sequenced* condition was similar in that the responses occurred at the same times as in the simultaneous case, but the stimulus was not played at these times. There was a three second pause after each trial. The trials of a given condition were grouped together, and participants were given visual prompts so that they knew when to respond. Each of the 75 trials within a condition used a different stimulus, taken from one of the five stimulus types described in Section 2.2, and presented in a random order. The two conditions were also presented in a random order.

2.2 Stimuli

Unlike previous imitation experiments which have used fixed-pitch stimuli, our experimental stimuli were synthesised from time-varying pitch trajectories in order to provide controlled conditions for testing the effect of specific deviations from constant pitch. Five stimulus types were chosen, representing a simplified model of the components of sung tones (constant pitch, initial and final glides, vibrato and pitch ramps). The pitch trajectories of the stimuli were generated from the models described below and synthesised by a custom-made MATLAB program, using a monotone male voice on the vowel /a:/.

The five different stimulus types considered in this work are: constant pitch (*stable*), a constant pitch preceded by an initial quadratic pitch glide (*head*), a constant pitch followed by a final quadratic pitch glide (*tail*), a linear pitch ramp (*ramp*), and a pitch with sinusoidal vibrato (*vibrato*). The stimuli are parametrised by the following variables: p_m, the main or central pitch; d, the duration of the transient part of the stimulus; and p_D, the extent of pitch deviation from p_m. For *vibrato* stimuli, d represents the period of vibrato. Values for each of the parameters are given in Table 1 and the text below.

By assuming an equal tempered scale with reference pitch A4 tuned to 440 Hz, pitch p and fundamental frequency f_0 can be related as follows [11]:

$$p = 69 + 12 \log_2 \frac{f_0}{440} \tag{1}$$

such that for integer values of p the scale coincides with the MIDI standard. Note that pitch is not constrained to integer values in this representation.

For the *stable* stimulus, the pitch trajectory $p(t)$ is defined as follows:

$$p(t) = p_m, \qquad 0 \leqslant t \leqslant 1. \tag{2}$$

The *head* stimulus is represented piecewise by a quadratic formula and a constant:

$$p(t) = \begin{cases} at^2 + bt + c, & 0 \leqslant t \leqslant d \\ p_m, & d < t \leqslant 1. \end{cases} \tag{3}$$

The parameters a, b and c are selected to make the curve pass through the point $(0, p_m + p_D)$ and have its vertex at (d, p_m). The *tail* stimulus is similar, with $p(t) = p_m$ for $t < 1 - d$, and the transient section being defined for $1 - d \leqslant t \leqslant 1$. In this case the parameters a, b and c are chosen so that the curve has vertex $(1 - d, p_m)$ and passes through the point $(1, p_m + p_D)$.

The *ramp* stimuli are defined by:

$$p(t) = p_m + p_D \times (t - 0.5), \qquad 0 \leqslant t \leqslant 1. \tag{4}$$

Finally, the equation of *vibrato* stimuli is:

$$p(t) = p_m + p_D \sin\left(\frac{2\pi t}{d}\right), \qquad 0 \leqslant t \leqslant 1. \tag{5}$$

There is a substantial amount of data on the fundamental frequency of the voice in the speech of speakers who differ in age and sex [23]. We chose three pitch values according to gender to fall within a comfortable range for most singers. The pitches C3 ($p = 48$), F3 ($p = 53$) and B♭3

($p = 58$) were chosen for male singers and C4 ($p = 60$), F4 ($p = 65$) and B♭4 ($p = 70$) for female singers. For the *vibrato* stimuli, we set the vibrato rate according to a reported mean vibrato rate across singers of 6.1 Hz [18], and the extent or depth of vibrato to ±0.25 or 0.5 semitones, in accordance with values reported by [21]. Because intonation accuracy is affected by the duration of the note [4, 6], we used a fixed one-second duration for all stimuli in this experiment.

Table 1: Parameter settings for each stimulus type. The octave for the pitch parameter was dependent on sex (3 for male, 4 for female).

Type	p_m	d	p_D	Count
stable	{C, F, B♭}	{0.0}	{0.0}	3
head	{C, F, B♭}	{0.1, 0.2}	{±1, ±2}	24
tail	{C, F, B♭}	{0.1, 0.2}	{±1, ±2}	24
ramp	{C, F, B♭}	{1.0}	{±1, ±2}	12
vibrato	{C, F, B♭}	{±0.32}	{0.25, 0.5}	12

2.3 Participants

A total of 43 participants (27 female, 16 male) took part in the experiment. 38 of them were recorded in the studio and 5 were distance participants from the USA, Germany, Greece and China (2 participants). The range of ages was from 19 to 34 years old (mean: 25.1; median: 25; std.dev.: 2.7). Apart from 3 participants who did not complete the experiment, most singers recorded all the trials.

We intentionally chose non-poor singers as our research target. "Poor-pitch singers" are defined as those who have a deficit in the use of pitch during singing [15, 25], and are thus unable to perform the experimental task. Participants whose pitch imitations had on average at least one semitone absolute error were categorised as poor-pitch singers. The data of poor-pitch singers is not included in this study, apart from one singer who occasionally sang one octave higher than the target pitch.

Vocal training is an important factor for enhancing the singing voice and making the singer's voice different from that of an untrained person [12]. To allow us to test for the effects of training, participants completed a questionnaire containing 34 questions from the Goldsmiths Musical Sophistication Index [13] which can be grouped into 4 main factors for analysis: active engagement, perceptual abilities, musical training and singing ability (9, 9, 7 and 7 questions respectively).

2.4 Recording Procedure

A tutorial video was played before participation. In the video, participants were asked to repeat the stimulus precisely. They were not told the nature of the stimuli. Singers who said they could not imitate the time-varying pitch trajectory were told to sing a stable note of the same pitch.

The experimental task consisted of 2 conditions, each containing 75 trials, in which participants sang three one-second responses in a 16-second period. It took just over one hour for participants to finish the experiment. 22 singers took the simultaneous condition first and 21 singers took the sequenced condition first. Although the synthetic stimulus simulated the vowel /a:/, participants occasionally chose other vowels that felt comfortable.

We used an on-line system to record and manage the experiment. After sign-up, participants completed the unfinished tests guided by a graphical interface. After singing each trial, the system automatically uploaded the recordings to a server and the annotation results were simultaneously generated. All responses were labelled with singer ID, condition, trial, order and repetition.

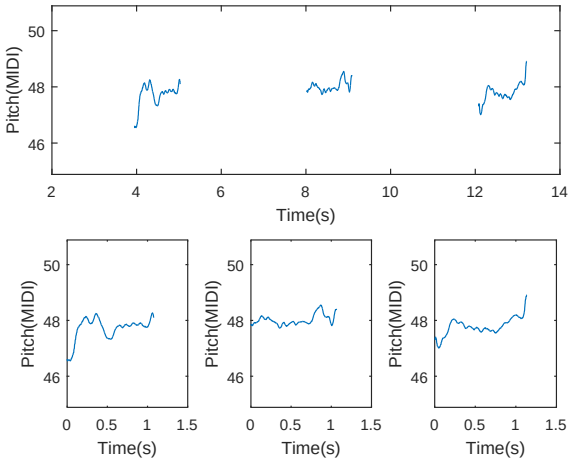

Figure 2: Example of extracted pitch and annotation for *head* stimulus ($p_m = 48$, $p_D = 1$, $d = 0.1$). The upper panel shows the results for pitch extraction by YIN, and the three lower panels show the segmented responses.

2.5 Annotation

Each recording file contains three responses, from which we extract pitch information using the YIN algorithm (version 28th July 2003) [5]. This outputs the pitch trajectory $p(t)$ from which we compute the median pitch $\bar{p}$ for each response. The segmentation into individual responses is based on the timing, pitch and power. If participants sang more than 3 repetitions we choose the three responses that have the longest duration and label them with the recording order. Any notes having a duration less than 0.1 seconds were excluded. Any remaining notes with a duration less than 0.4 seconds were flagged and checked manually. Most of these deficient notes were due to participants making no response. Figure 2 shows an example of pitch extraction and segmentation.

The main pitch $\bar{p}$ of response was calculated by removing the first 10% and last 10% of the response duration, and computing the median of the remaining pitch track. The pitch error e^p is calculated as the difference between the main pitch of the stimulus p_m and that of the response $\bar{p}$:

$$e^p = \bar{p} - p_m \tag{6}$$

For avoiding bias due to large errors we exclude any responses with $|e^p| > 2$ (4% of responses). Such errors arose

when participants sang the pitch of the previous stimulus or one octave higher than the stimulus. The resulting database contains 18572 notes, from which the statistics below were calculated.

The mean pitch error (MPE) over a number of trials measures the tendency to sing sharp (MPE > 0) or flat (MPE < 0) relative to the stimulus. The mean absolute pitch error (MAPE) measures the spread of a set of responses. These can be viewed respectively as inverse measures of accuracy and precision (cf. [16]).

To analyse differences between the stimulus and response as time series, pitch error $e^p_f(t)$ is calculated frame-wise: $e^p_f(t) = p_r(t) - p_s(t)$, for stimulus $p_s(t)$ and response $p_r(t)$, where the subscript f distinguishes frame-wise results. For frame period T and frame index i, $0 \leqslant i < M$, we calculate summary statistics:

$$\mathrm{MAPE}_f = \frac{1}{M} \sum_{i=0}^{M-1} |e^p_f(iT)| \tag{7}$$

and MPE_f is calculated similarly. Equation 7 assumes that the two sequences $p_r(t)$ and $p_s(t)$ are time-aligned. Although cross-correlation could be used to find a fixed offset between the sequences, or dynamic time warping could align corresponding features if the sequences proceed at different or time-varying rates, in our case we consider singing with the correct timing to be part of the imitation task, and we align the stimulus to the beginning of the detected response.

3. RESULTS

We first report pitch error (MPE: 0.0123; std.dev.: 0.3374), absolute pitch error (MAPE: 0.2441; std.dev.: 0.2332) and frame-wise absolute pitch error (MAPE_f: 0.3968; std.dev.: 0.2238) between all the stimuli and responses. 71.1% of responses have an absolute error less than 0.3 semitones. 51.3% of responses are higher than the stimulus ($e^p > 0$). All the singers' information, questionnaire responses, stimulus parameters and calculated errors were arranged in a single table for further processing. We first analyse the factors influencing absolute pitch error in the next two subsections, and then consider pitch error in section 3.3 and the modelling of responses in the following two subsections.

3.1 Influence of stimulus type on absolute pitch error

We performed one-way independent samples analysis of variance (one-way ANOVA) with the fixed factor stimulus type (five levels: *stable*, *head*, *tail*, *ramp* and *vibrato*) and the random factor participant. There was a significant effect of stimulus type ($[F(4, 18567) = 72.3, p < .001]$).

Post hoc comparisons using the Tukey HSD test indicated that the absolute e^p for *tail*, *ramp* and *vibrato* stimuli were significantly different from that of the *stable* stimuli, while the *head* stimuli showed no significant difference from *stable* stimuli (see Table 2). Thus *tail*, *ramp* and *vibrato* stimuli do have an effect on pitch precision. Table 2 also shows the 95% confidence intervals for each stimulus type. Effect sizes were calculated by a linear mixed-effects model comparing with *stable* stimulus results.

Stimulus	MAPE	Confidence interval	Effect size
stable	0.1977	[0.1812, 0.2141]	–
head	0.1996	[0.1938, 0.2054]	0.2 cents
tail	0.2383	[0.2325, 0.2441]*	4.1 cents
ramp	0.3489	[0.3407, 0.3571]***	15.1 cents
vibrato	0.2521	[0.2439, 0.2603]***	5.5 cents

Table 2: Mean absolute pitch error (MAPE) and 95% confidence intervals for each stimulus type (***p $<$ .001; **p $<$ .01; *p $<$ .05).

3.2 Factors of influence for absolute pitch error

The participants performed a self-assessment of their musical background with questions from the Goldsmiths Musical Sophistication Index [14] covering the four areas listed in Table 3, where the general factor is the sum of other four factors. An ANOVA F-test found that all background factors are significant for pitch accuracy (see Table 3). The task involved both perception and production, so it is to be expected that both of these factors (perceptual and singing abilities) would influence results. Likewise most musical training includes some ear training which would be beneficial for this experiment.

Factor	Test Results
General factor	$F(30, 18541) = 54.4$ ***
Active engagement	$F(21, 18550) = 37.3$ ***
Perceptual abilities	$F(22, 18549) = 57.5$ ***
Musical training	$F(24, 18547) = 47.2$ ***
Singing ability	$F(20, 18551) = 69.8$ ***

Table 3: Influence of background factors.

We used R [19] and *lme4* [2] to perform a linear mixed-effects analysis of the relationship between factors of influence and $|e^p|$. The factors stimulus type, main pitch, age, gender, the order of stimuli, trial condition, repetition, duration of pitch deviation d, extent of pitch deviation p_D, observed duration and the four factors describing musical background were added separately into the model, and a one-way ANOVA between the model with and without the factor tested whether the factor had a significant effect. Table 4 shows the p-value of ANOVA results after adding each factor.

We created a fixed model with factors stimulus type, main pitch, repetition and trial condition. As a random effect, we had the factor of the singer. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. The p-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question [26].

According to the modelling results on $|e^p|$, significant effects were found for the factors stimulus type, main pitch

Table 4: Significance and effect sizes for tested factors based on ANOVA results.

Factors	p-value	Effect size (cents)
Stimulus type	2.2e-16***	See Table 2
p_m	5.4e-7***	-0.19
Age	0.51	
Gender	0.56	
Order of stimuli	0.13	
Trial condition	2.2e-16***	3.2
Repetition	2.2e-16***	-1.8
Duration of transient d	2.2e-16***	11.4
sign(p_D)	5.1e-6***	0.8
abs(p_D)	8.3e-12***	1.9
Observed duration	3.3e-4***	-5.4
Active engagement	6.9e-2	
Perceptual abilities	0.04*	-0.3
Musical training	6.2e-5***	-0.5
Singing ability	8.2e-2	

p_m (effect size: -2.35 cents per octave), trial condition, repetition, musical background, duration of pitch deviation (effect size: 11.4 cents per second), direction of pitch deviation, magnitude of pitch deviation (effect size: 1.7 cents per semitone) and observed duration (effect size: -5.4 cents per second). The remaining factors (singer, age, gender and the order of stimuli) did not have any significant effect on $|e^p|$ in this model. The LME models gave different results for the background questionnaire factors than the one-way ANOVA, with only two of the factors, perceptual abilities and musical training, having a significant effect.

Contrary to our hypothesis, singing simultaneously (MAPE: 0.26; std.dev.: 0.25) is 3.2 cents less accurate than the sequenced condition (MAPE: 0.23; std.dev: 0.21). Despite the large spread of results, the standard errors in the means are small and the difference is significant. Recall also that responses with $|e^p|$ over 2 semitones were excluded.

Other significant factors were repetition, where we found that MAPE decreases 1.8 cents for each repetition (that is, participants improved with practice), and observed duration and main pitch, which although significant, had very small effect sizes for the range of values they took on.

3.3 Effect of pitch deviation on pitch error

We now look at specific effects on the direction of pitch error, to test the hypothesis that asymmetric deviations from main pitch are likely to lead to errors in the direction of the deviation. For the *stable*, *head* and *tail* stimuli, a correlation analysis was conducted to examine the relationship between pitch deviation and MPE. The result was significant on MPE ($F(4, 12642) = 8.4, p = 9.6e{-}7$) and MAPE ($F(4, 12642) = 8.2, p = 1.3e{-}6$). A significant regression equation was found, with $R^2 = 2.5e-3$, modelling pitch error as $e^P = 0.033 + 0.01 p_D$. Pitch error increased 1 cent for each semitone of p_D, a significant but small effect, as shown in Figure 3.

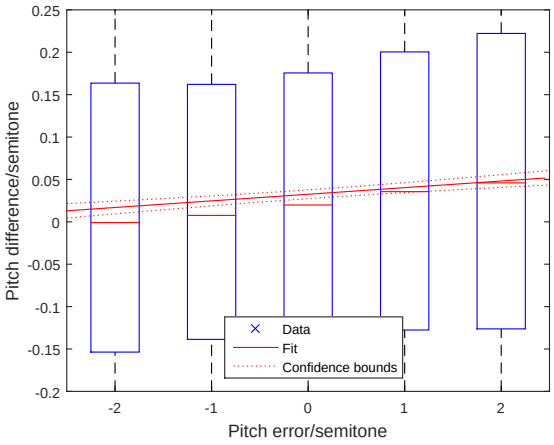

Figure 3: Boxplot of MPE for different p_D, showing median and interquartile range, regression line (red, solid) and 95% confidence bounds (red, dotted). The regression shows a small bias due to the positively skewed distribution of MPE.

3.4 Modelling

In this section, we fit the observed pitch trajectories to a model defined by the stimulus type, to better understand how participants imitated the time-varying stimuli. The *head* and *tail* stimuli are modelled by a piecewise linear and quadratic function. Given the break point, corresponding to the duration of the transient, the two parts can be estimated by regression. We perform a grid search on the break point and select the optimal parameters according to the smallest mean square error. Figure 4 shows an example of *head* response modelling.

The *ramp* response is modelled by linear regression. The model p_m of a *stable* response is the median of $p(t)$ for the middle 80% of the response duration. The *vibrato* responses were modelled with the MATLAB nlinfit function using Equation 5 and initialising the parameters with the parameters of the stimulus.

For the absolute pitch error between modelling results and stimuli, 66.5% of responses have an absolute error less than 0.3 semitones, while only 29.3% of trials have an absolute error less than 0.3 semitones between response and stimulus. We observed that some of the *vibrato* models did not fit the stimulus very well because the singer attempted to sing a stable pitch rather than imitate the intonation trajectory.

3.5 Duration of transient

As predicted, the duration d of the transient has a significant effect on MPE ($F(5, 18566) = 51.4, p < .001$). For the *stable*, *head* and *tail* stimuli, duration of transient influences MAPE ($F(2, 12644) = 31.5, p < .001$), where stimuli with smaller transient length result in lower MAPE. The regression equation is $MAPE = 0.33 + 0.23d$ with $R^2 = 0.208$. MAPE increased 23.2 cents for each second of transient. This matches the result from the linear mixed-effects model, where effect size is 23.8 cents per second.

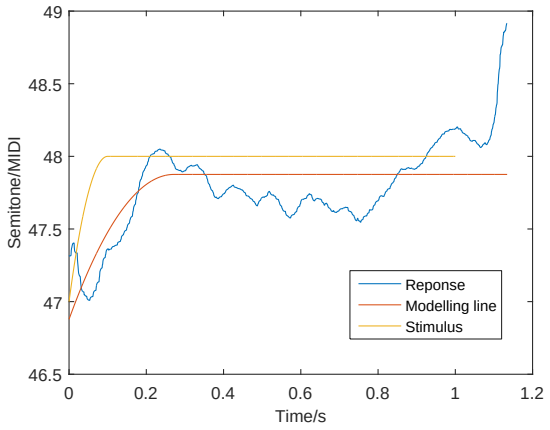

Figure 4: Example of modelling the response to a *head* stimulus with parameters $\mathrm{d} = 0.1$, $\mathrm{p_D} = -1$ and $\mathrm{p_m} = 48$. The response model has $\mathrm{d} = 0.24$, $\mathrm{p_D} = -0.997$ and $\mathrm{p_m} = 47.87$. The forced fit to the stimulus model treats as noise response features such as the final rising intonation.

Based on the modelling results, we observed that transient length in responses was longer than in the corresponding stimuli. 74.2% of *head* and *tail* responses have transient length longer than that of the stimulus. Stimulus transients are 0.1 or 0.2 seconds, but 65.5% of *head* and 72.0% of *tail* responses have a transient longer than 0.2 seconds.

4. DISCUSSION

Since we intentionally chose non-poor singers, most participants imitated with small error. 88.5% of responses were sung with intonation error less than half a semitone. The responses are characterised far more by imprecision than inaccuracy. That is, there is very little systematic error in the results (MPE $= 0.0123$), whereas the individual responses exhibit much larger errors in median pitch (MAPE $= 0.2441$) and on a frame-wise level within notes ($\mathrm{MAPE_f} = 0.3968$). The results for MAPE are within the range reported for non-poor singers attempting known melodies (19 cents [11], 28 cents [4]), and thus is better explained by limitations in production and perception rather than by any particular difficulty of the experimental task.

The *stable* stimuli gave rise to the lowest pitch errors, although the *head* responses were not significantly different. The larger errors observed for the *tail*, *ramp* and *vibrato* stimuli could be due to a memory effect. These three stimulus types have in common that the pitch at the end of the stimulus differs from $\mathrm{p_M}$. Thus the most recent pitch heard by the participant could distract them from the main target pitch. The *ramp* stimuli, having no constant or central pitch, was the most difficult to imitate, and resulted in the highest MAPE.

It was hypothesised that the simultaneous condition would be easier than the sequenced condition, as singing tends to be more accurate when accompanied by other singers or instruments. We propose two reasons why this experiment might be exceptional. Firstly, in the sequenced condition, the time between stimulus and response was short (1 second), so it would be unlikely that the participant would forget the reference pitch. Secondly, the stimulus varied more quickly than the auditory feedback loop, the time from perception to a change in production (around 100ms [3]), could accommodate. Thus the feedback acts as a distractor rather than an aid. Singing in practice requires staying in tune with other singers and instruments. If a singer takes their reference from notes with large pitch fluctuations, especially at their ends, this will adversely affect intonation.

5. CONCLUSIONS

We designed a novel experiment to test how singers respond to controlled stimuli containing time-varying pitches. 43 singers vocally imitated 75 instances of five stimulus types in two conditions. It was found that time-varying stimuli are more difficult to imitate than constant pitches, as measured by absolute pitch error. In particular, stimuli which end on a pitch other than the main pitch (*tail*, *ramp* and *vibrato* stimuli) had significantly higher absolute pitch errors than the *stable* stimuli, with effect sizes ranging from 15 cents (*ramp*) to 4.1 cents (*tail*).

Using a linear mixed-effects model, we determined that the following factors influence absolute pitch error: stimulus type, main pitch, trial condition, repetition, duration of transient, direction and magnitude of pitch deviation, observed duration, and self-reported musical training and perceptual abilities. The remaining factors that were tested had no significant effect, including self-reported singing ability, contrary to other studies [11].

Using one-way ANOVA and linear regression, we found a positive correlation between extent of pitch deviation (pitch difference, $\mathrm{p_D}$) and pitch error. Although the effect size was small, it was significant and of similar order to the overall mean pitch error. Likewise we observed that the duration d of the transient proportion of the stimulus correlated with absolute pitch error. Contrary to expectations, participants performed 3.2 cents worse in the condition when they sang simultaneously with the stimulus, although they also heard the stimulus between singing attempts, as in the sequenced condition.

Finally, we extracted parameters of the responses by a forced fit to a model of the stimulus type, in order to describe the observed pitch trajectories. The resulting parameters matched the stimuli more closely than the raw data did. Many aspects of the data remain to be explored, but we hope that the current results take us one step closer to understanding interaction between singers.

6. DATA AVAILABILITY

There is the tutorial video which show participants how to finish the experiment before they start: `https://www.youtube.com/watch?v=xadECsaglHk`. The annotated data and code to reproduce our results are available in an open repository at: `https://code.soundsoftware.ac.uk/projects/stimulus-intonation/repository`.

7. REFERENCES

[1] Per-Gunnar Alldahl. *Choral Intonation*. Gehrmans, 2008.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[3] T. A. Burnett, M. B. Freedland, C. R. Larson, and T. C. Hain. Voice F0 Responses to Manipulations in Pitch Feedback. *Journal of the Acoustical Society of America*, 103(6):3153–3161, 1998.

[4] Jiajie Dai, Matthias Mauch, and Simon Dixon. Analysis of Intonation Trajectories in Solo Singing. In *Proceedings of the 16th ISMIR Conference*, volume 421, 2015.

[5] Alain De Cheveigné and Hideki Kawahara. YIN, A Fundamental Frequency Estimator for Speech and Music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[6] J. Fyk. Pitch-matching Ability In Children As A Function of Sound Duration. *Bulletin of the Council for Research in Music Education*, pages 76–89, 1985.

[7] David Gerhard. Pitch Track Target Deviation in Natural Singing. In *ISMIR*, pages 514–519, 2005.

[8] Joyce Bourne Kennedy and Michael Kennedy. *The Concise Oxford Dictionary of Music*. Oxford University Press, 2004.

[9] Peggy A Long. Relationships Between Pitch Memory in Short Melodies and Selected Factors. *Journal of Research in Music Education*, 25(4):272–282, 1977.

[10] Matthias Mauch and Simon Dixon. pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014.

[11] Matthias Mauch, Klaus Frieler, and Simon Dixon. Intonation in Unaccompanied Singing: Accuracy, Drift, and a Model of Reference Pitch Memory. *The Journal of the Acoustical Society of America*, 136(1):401–411, 2014.

[12] Ana P Mendes, Howard B Rothman, Christine Sapienza, and WS Brown. Effects of Vocal Training on the Acoustic Parameters of the Singing Voice. *Journal of Voice*, 17(4):529–543, 2003.

[13] Daniel Müllensiefen, Bruno Gingras, Jason Musil, Lauren Stewart, et al. The Musicality of Non-musicians: An Index for Assessing Musical Sophistication in the General Population. *PloS one*, 9(2):e89642, 2014.

[14] Daniel Müllensiefen, Bruno Gingras, Lauren Stewart, and J Musil. The Goldsmiths Musical Sophistication Index (Gold-MSI): Technical Report and Documentation v0.9. *London: Goldsmiths, University of London. URL: http://www.gold.ac.uk/music-mind-brain/gold-msi*, 2011.

[15] Peter Q Pfordresher and Steven Brown. Poor-pitch Singing in the Absence of "Tone Deafness". *Music Perception: An Interdisciplinary Journal*, 25(2):95–115, 2007.

[16] Peter Q Pfordresher, Steven Brown, Kimberly M Meier, Michel Belyk, and Mario Liotti. Imprecise Singing is Widespread. *The Journal of the Acoustical Society of America*, 128(4):2182–2190, 2010.

[17] John Potter, editor. *The Cambridge Companion to Singing*. Cambridge University Press, 2000.

[18] Eric Prame. Measurements of the Vibrato Rate of Ten Singers. *The Journal of the Acoustical Society of America*, 96(4):1979–1984, 1994.

[19] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[20] John Andrew Simpson, Edmund S.C. Weiner, et al. *The Oxford English Dictionary*, volume 2. Clarendon Press Oxford, 1989.

[21] J. Sundberg. Acoustic and Psychoacoustic Aspects of Vocal Vibrato. Technical Report STL-QPSR 35 (2–3), pages 45–68, Department for Speech, Music and Hearing, KTH, 1994.

[22] Annie H Takeuchi and Stewart H Hulse. Absolute Pitch. *Psychological bulletin*, 113(2):345, 1993.

[23] Hartmut Traunmüller and Anders Eriksson. The Frequency Range of the Voice Fundamental in the Speech of Male and Female Adults. *Consulté le*, 12(02):2013, 1995.

[24] Allan Vurma and Jaan Ross. Production and Perception of Musical Intervals. *Music Perception: An Interdisciplinary Journal*, 23(4):331–344, 2006.

[25] Graham F Welch. Poor Pitch Singing: A Review of the Literature. *Psychology of Music*, 7(1):50–58, 1979.

[26] Bodo Winter. Linear Models and Linear Mixed Effects Models in R with Linguistic Applications. *arXiv preprint arXiv:1308.5499*, 2013.

[27] Yi Xu and Xuejing Sun. How Fast Can We Really Change Pitch? Maximum Speed of Pitch Change Revisited. In *INTERSPEECH*, pages 666–669, 2000.

[28] Jean Mary Zarate and Robert J Zatorre. Experience-dependent Neural Substrates Involved in Vocal Pitch Regulation During Singing. *Neuroimage*, 40(4):1871–1887, 2008.