Perceptual evaluations of voice scatter in unison choir sounds

Ternström, S.

journal: STL-QPSR
volume: 32
number: 2-3
year: 1991
pages: 041-049

http://www.speech.kth.se/qpsr
PERCEPTUAL EVALUATIONS OF VOICE SCATTER IN UNISON CHOIR SOUNDS*

Sten Ternström

Abstract

The preferences of experienced listeners for pitch and formant frequency dispersion in unison choir sounds were explored using synthesized stimuli. Two types of dispersion were investigated: 1) pitch scatter, which arises when different voices in an ensemble exhibit small differences in mean fundamental frequency; and 2) spectral smear, here defined as such dispersion of formants 3 to 5 as arises when the singers have vocal tracts of different lengths. Each stimulus represented a choir section of five bass, tenor, alto, or soprano voices, producing the vowels [u], [a], or [æ]. Subjects chose one dispersion level out of six available, selecting the "maximum tolerable" in a first run and the "preferred" in a second run.

The listeners were very different in their tolerance for dispersion. Typical scatter choices (distribution modes) were 14 cent standard deviation for "tolerable" and 0 or 5 cent for "preferred". The smear choices were less consistent, averages being 12% standard deviation for "tolerable" and 7% for "preferred". In all modes of assessment, the highest levels of dispersion were chosen for the vowel [u] on a bass tone. There was a vowel effect on the smear choices. The effects of voice category were not significant.

INTRODUCTION

Ensemble sounds

While a single performer usually produces tones of well-defined pitch, loudness, and timbre, an ensemble of performers will generate sounds characterized by statistical distributions of these properties. For example, the sound of an ensemble playing in unison does not evoke in the listener a number of closely adjacent pitches, but rather a single, albeit perhaps less well-defined pitch, and possibly also a sensation of pitch dispersion. The aim of the present investigation was to determine the degree of dispersion in fundamental frequency ($F_0$) and in formant frequencies F3 to F5 that would be preferred and/or tolerated by experts listening to unison choir sounds. A further objective was to determine whether such preferences/tolerances are affected by vowel timbre and $F_0$ level.

It is not clear how the above types of dispersion among singer voices would affect the statistical properties of the total sound, since each singer has a personal voice flutter. By flutter, we mean small variations in $F_0$ that are too rapid to be perceived as pitch variations, yet too slow to affect timbre; i.e., in the approximate range 5-15 Hz (Ternström & Friberg, 1989). We, therefore, opted not to evaluate scatter from recordings of live choir sounds. With recent advances in singing synthesis, it is now practical instead to create choral timbres from synthesized voices, with precisely known characteristics.

Fundamental frequency dispersion - scatter

To our knowledge, very few researchers have reported measurements of $F_0$ dispersion in choirs or other ensembles. Lottermoser & Meyer (1960) used commercial recordings of three reputable choirs to study choral intonation. The dispersion in fundamental frequency, measured as the bandwidth of partial tones, varied greatly in these choirs, from 2 to 60 cent, but was typically in the

*This paper was presented at the 20th Symp.: Care of the Professional Voice, Philadelphia, PA, 1991.
range 20-30 cent. These measurements would include the effects of flutter, and so are difficult to interpret in terms of musical intonation. If instead the $F_0$ of each singer is averaged over the duration of each tone, this average ($MF_0$) is relevant to intonation. In an earlier paper (Lottermoser & Meyer, 1960), scatter was defined as the standard deviation over voices in $MF_0$. This measure reflects only the discrepancies in mean $F_0$ between singers, and is insensitive to the amount of flutter in the voices. Ternström & Sundberg (1988) made multitrack recordings of choir basses in rehearsal, and found the scatter to be 10-15 cents, in trials judged by their conductor to be musically acceptable. Dolson (1983) reports making satisfactory simulations of violin ensemble when the individual pitches were chosen at random from a range of ±1.5% (about ±25 cents).

Flutter and scatter are not entirely independent entities. Random flutter will necessarily give rise to fluctuating scatter, even if the mean fundamental frequency is the same for all voices, simply because the voices are likely to be spaced apart in $F_0$ at any given moment. Informal experiments in choir synthesis (Ternström, Friberg, & Sundberg, 1988) show that increasing only the flutter often conveys an impression of a larger choir. The converse is not true, however. Scatter alone does not sound like flutter, but rather like poor tuning.

It is possible that the acoustical differences between the voices of choir basses, tenors, altos, and sopranos lead to different flutter and/or scatter behavior in the different sections of the choir. For example, it is conceivable that sopranos would need to sing in closer unison than do basses. The reason for this would be twofold. Firstly, the high $F_0$ of sopranos causes more rapid beating at a given ratio of mistuning. The beat frequency for a mistuning of 5 cents will be 2.5 Hz when $F_0 = 880$ Hz, a high soprano note, but only 0.32 Hz at $F_0 = 110$ Hz, a typical bass note. Secondly, the soprano beats would be far more intense, because the energy of high soprano tones is concentrated to the fundamental partial; this is seldom the case for bass tones. We therefore wished to test the hypothesis that pitch scatter is not equally preferred or tolerated for different values of $F_0$. The frequencies chosen were 110, 220, 330, and 623 Hz, being typical of the four choir sections.

Vowel timbre, too, might affect the acceptability of pitch scatter. Informal trials with synthesis indicated that a large scatter could be less objectionable on a closed vowel such as [u] than on an open [a] or [æ] vowel. A second hypothesis was, therefore, that pitch scatter is not equally preferred or tolerated for all vowels.

Formant frequency dispersion - spectral smear

Many choir conductors contend that a uniform pronunciation of vowels is beneficial to the overall timbre and sometimes also to intonation. The vowel identity is determined mainly by the frequencies $F_1$ and $F_2$ of the two lowest formants. In an earlier study (Ternström & Sundberg, 1989), it was found that bass singers who were long-time members of one choir were more agreed on $F_1$ and $F_2$ when singing alone than when speaking alone. This might indicate that they had adopted a sort of "choral dialect". Typical standard deviations were 5-10% in the frequencies of formants $F_1$-$F_4$, taken across singers. The same amount of variation was found in a systematic inventory of the vowel articulation of choir singers (Carlsson & Ternström, forthcoming). In preliminary synthesis trials, however, we were surprised to find that the subjective effect of dispersing $F_1$ and $F_2$ across voices was quite small; too small in fact, for us to feel confident about making a listening test.

Another issue relevant to choral timbre is the blending of voices such that the ensemble sound is homogeneous and that no singer is heard to "stand out." The vocal personality of a singer's timbre is due largely to the higher formants $F_3$, $F_4$ and $F_5$, whose frequencies are directly related to vocal tract length but show little dependence on vowel. For simplicity, we decided to treat these formants as a cluster, and to simulate a diversity of singer physiognomies by scaling this cluster up or down in frequency. The standard deviation over voices of the scaling factor will be called smear, for the remainder of this paper. As an example, five voices with the cluster of higher formants transposed in frequency by $-7$, $-4$, 0, $+4$, and $+7\%$ respectively will be said to have a smear of 5%. In synthesis trials, smear thus defined was easier to perceive than had been a similar amount of dispersion in $F_1$ and $F_2$. 
Preferences and tolerances for smear might also be affected by $F_0$ level and vowel. At high fundamental frequencies, the partial tones of soprano sounds are sparse and should convey little information about smear. Low bass notes, on the other hand, potentially supply a great profusion of partials, even in the frequency range of the higher formants (2-4 kHz). One might also expect an effect of vowel: the closed vowel [u] lacks energy in this range - at least in choir singing, without the 'singer's formant' - whereas [æ] and [a] do not. For these reasons, we also wished to test the hypotheses that the preferred and tolerated levels of smear depend on $F_0$ level and vowel timbre.

Choice of "normal" voice

While the above considerations suggest a strategy for creating dispersion in voice sounds, one must still decide on the parameters to use for synthesis of the "normal" or center voice. Normally, to test the effect of $F_0$ changes, we would keep other voice parameters, such as $F_1$, $F_2$, and the flutter level, constant across $F_0$ values. However, this would result in many synthesized vowels sounding entirely unnatural, because $F_1$ and $F_2$ normally vary somewhat with $F_0$ and with the sex of the singer. Similarly, it would be inappropriate to equalize for loudness, since different vowels and $F_0$ values normally yield different loudness levels for the same singer effort. We also found that a flutter level which sounds reasonable for a bass would sound far too wobbly for a soprano voice. (This suggests a future investigation into the subjective preferences for flutter.)

For the "male" voices, the formant frequencies were chosen as averages of live singers measured in Ternström & Sundberg (1989). At the time of the present investigation, formant frequencies of female choir singers had not yet been analyzed, so female choir singers were synthesized using formant values chosen from a combination of personal research experience and subjective impressions. The latter applied also to the choice of flutter levels.

METHOD

Synthesis of stimuli

A newly developed numerical synthesizer of singing voice (Carlsson, Ternström, Sundberg, & Ungvary, 1991) was used to generate data files containing sampled waveforms of individual voices. The synthesizer is implemented both in a digital signal processor as an interactive, real-time device, and in a personal computer as a non-interactive, batch-mode program that produces a data file with the sampled waveform. The interactive mode is suitable for rapid testing of formant configurations, flutter, and other parameters of individual voices. The batch mode is convenient for generating sequences of systematically varied stimuli.

It was found that a stereophonic presentation was much more convincing than a monophonic one. Five synthesized voices were combined into a stereophonic ensemble, by adding them together with different time delays for the left and right channels. Such binaural presentation is preferable to normal amplitude-based "panning", for two reasons. One is that the stereo image over headphones is more realistic. The other is that all five voices are presented at the same intensity to both ears, which eliminates such perceptual bias as might arise from asymmetries in headphone position or in subject hearing loss.

Two series of stimulus sounds were synthesized, one for scatter and the other for smear. There were twelve conditions in each series, combining three vowels with four voice categories (Table I). Each condition was synthesized at six levels of dispersion (Table II). The corpus of 144 stereophonic tokens ($2 \times 12 \times 6$) was compiled into a bank of sampled data files using signal editing software (Soundswell Music Acoustics).

Procedure

Each subject performed four trials, in which the presented sounds were to be adjusted for 1) maximum tolerated scatter; 2) maximum tolerated smear; 3) preferred scatter, and 4) preferred smear. Each subject was given a short dummy run to clarify the procedure. Before each trial, the subject was orally instructed as to the new task. The experiment was conducted in a quiet office environment. The subjects sat in front of a computer, wearing light, open-speaker headphones (Sennheiser...
HD414SL), and using only a mouse to control the presentation.

Table I: Synthesis parameters used for generating the "center" voices. Formant frequencies $F_n$ and bandwidths $B_n$ are given in Hz.

<table>
<thead>
<tr>
<th>Voice category</th>
<th>Vowel</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_3-F_5$</th>
<th>$F_4$</th>
<th>$F_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOPRANO</td>
<td>[u]</td>
<td>664</td>
<td>45</td>
<td>1120</td>
<td>65</td>
<td>3000</td>
<td>170</td>
</tr>
<tr>
<td><strong>F₀</strong>: Eb5, 623 Hz</td>
<td>[a]</td>
<td>660</td>
<td>40</td>
<td>1320</td>
<td>65</td>
<td>3150</td>
<td>200</td>
</tr>
<tr>
<td>Flutter level 5 cent</td>
<td>[æ]</td>
<td>662</td>
<td>50</td>
<td>1450</td>
<td>70</td>
<td>2850</td>
<td>200</td>
</tr>
<tr>
<td>ALTO</td>
<td>[u]</td>
<td>410</td>
<td>60</td>
<td>880</td>
<td>80</td>
<td>2900</td>
<td>129</td>
</tr>
<tr>
<td><strong>F₀</strong>: E3, 330 Hz</td>
<td>[æ]</td>
<td>700</td>
<td>50</td>
<td>1770</td>
<td>110</td>
<td>2510</td>
<td>180</td>
</tr>
<tr>
<td>Flutter level 9 cent</td>
<td>[a]</td>
<td>700</td>
<td>55</td>
<td>1100</td>
<td>80</td>
<td>3060</td>
<td>160</td>
</tr>
<tr>
<td>TENOR</td>
<td>[u]</td>
<td>320</td>
<td>70</td>
<td>720</td>
<td>100</td>
<td>2500</td>
<td>200</td>
</tr>
<tr>
<td><strong>F₀</strong>: A3, 220 Hz</td>
<td>[æ]</td>
<td>480</td>
<td>70</td>
<td>1650</td>
<td>100</td>
<td>2460</td>
<td>200</td>
</tr>
<tr>
<td>Flutter level 14 cent</td>
<td>[a]</td>
<td>550</td>
<td>70</td>
<td>1050</td>
<td>100</td>
<td>2750</td>
<td>200</td>
</tr>
<tr>
<td>BASS</td>
<td>[u]</td>
<td>300</td>
<td>70</td>
<td>630</td>
<td>100</td>
<td>2300</td>
<td>200</td>
</tr>
<tr>
<td><strong>F₀</strong>: A2, 110 Hz</td>
<td>[æ]</td>
<td>480</td>
<td>70</td>
<td>1600</td>
<td>100</td>
<td>2280</td>
<td>200</td>
</tr>
<tr>
<td>Flutter level 14 cent</td>
<td>[a]</td>
<td>520</td>
<td>70</td>
<td>900</td>
<td>100</td>
<td>2650</td>
<td>200</td>
</tr>
</tbody>
</table>

Table II: Standard deviations used for the six levels of dispersion.

<table>
<thead>
<tr>
<th>Dispersion level</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scatter in cents</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>14</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Smear in %</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td>14</td>
<td>20</td>
</tr>
</tbody>
</table>

The experimental procedure was automated.* Once the experimenter had initiated a series, the software would
- randomize the presentation order of the twelve vowel/voice conditions - choose an initial dispersion level at random, avoiding the extremes
- play back the first token
- present a dialog box to the subject (Fig. 1)
- let the subject repeatedly select a token with more or less dispersion and play this back
- record the subject's final choice and proceed to the next condition

*The experiment was controlled using the macro facilities of the Microsoft Excel spreadsheet program. A special playback macro was linked to a Soundswell device driver controlling an Ariel DSP-16 board with twin D/A outputs.
There was no time constraint. The subjects took between seven and fifteen minutes to complete one trial.

**Subjects**

There were nine subjects, with at least university level training in music, and all with extensive and advanced choral experience. Five subjects were active choir conductors, and four were long-time singers in elite Stockholm choirs. There were four males and five females, their ages ranging from 24 to 45 years.

---

**RESULTS**

**Validity**

Most subjects spontaneously made comments to the effect that the syntheses sounded realistic (see the Appendix). Although people with extensive choral experience presumably would be able to assess dispersion in pitch or timbre, some confirmation is needed that the task was well understood by the subjects. Figure 2 shows the total distribution of responses for scatter and smear. For scatter, the difference in the response distributions between "tolerated" and "preferred" is very clear, indicating that the subjects were indeed applying different criteria to the same stimuli. For smear, the difference is still clear but less pronounced. It seems reasonable to infer from these distributions that high levels of dispersion are generally considered undesirable, as would be expected. On the whole, then, we would expect a consistent subject to choose as "preferred" a somewhat lower level of dispersion than that chosen as "tolerated". The pairwise difference between a subject's "tolerated" and "preferred" choices in a given condition will be called the **tolerance**. The average tolerance of a subject may be large or small but is expected to be positive. Furthermore, if the average tolerance is small, then the variance of the tolerance should be small also. If these expectations are not met, the subject may have been choosing haphazardly, possibly due to a lack of clear preferences; or, he (or she) may have significantly shifted criterion during the course of one trial. The latter would be a kind of learning effect.
Fig. 3. Assessment of reliability, per subject. Column heights show the mean tolerance, i.e., the mean pairwise difference between twelve "tolerated" and "preferred" responses. Markers connected by a dashed line show the standard deviation of the tolerance. Scatter tolerances (top) and smear tolerances (bottom). For comments, see text.

Figure 3 shows the averages and standard deviations of the tolerances of the nine subjects. It is evident that they had very different latitudes in their judgments. For example, subject A appears to be strict, in that she tolerated little more scatter than she preferred, while subject C was more liberal, tolerating much more scatter than he would prefer. Nevertheless, for the scatter choices, all subjects exhibit positive averages of tolerance, and the standard deviations are mostly small. The smear tolerances are much more variable, and one subject (F) exhibits a negative mean with a large standard deviation. Our conclusion is that the scatter choices may be taken as relevant and informative, while the smear choices are weaker.

Factor effects
The results are shown in Figs. 2 and 4. An analysis of variance was performed to determine the significance of the effect of the factors "vowel" and "voice category" (Table I). For smear, the "vowel" factor was significant at the 0.90 level for both "tolerated" and "preferred", meaning that smear is less acceptable on some vowels than on others. This factor effect covaried with voice category; for male, but not for female voice categories, the highest smear choices were made for the vowel [u]. It may be noted also that the highest averages in all four trials resulted from the combination of bass-like voices with the vowel [u].

DISCUSSION
Limitations
The experimental factor "voice category" was a composite one, with the advantage of realism but with the disadvantage that several acoustic parameters changed from one voice category to the next (Table I). The factor "vowel", too, is composite in that several formant frequencies are changed from one vowel to another. The lack of significant effects of voice category may have been due to
conflicting interactions between more elementary acoustic parameters. For example, while tolerated scatter was highest on the vowel [u] for the voice categories bass, tenor, and alto, this vowel exhibited the lowest tolerated scatter for soprano voices. In the set of stimuli used here, soprano [u] sounds were special with regard to the extreme spectral dominance of the fundamental partial. This resulted in a larger amplitude modulation index of the total sound than for any other combination of factors. This may have had an effect on subject preferences.

![Graphs showing tolerated and preferred scatter and smear for different voice categories and vowels.](image)

**Fig. 4.** Response averages for scatter (top) and smear (bottom): tolerated (left), and preferred (right). Each value is an average over nine subjects. See also Table III.

<table>
<thead>
<tr>
<th>Rated quantity</th>
<th>Source of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scatter</td>
<td>Vowel</td>
</tr>
<tr>
<td>Tolerated</td>
<td>0.75</td>
</tr>
<tr>
<td>Preferred</td>
<td>0.75</td>
</tr>
<tr>
<td>Smear</td>
<td>Tolerated</td>
</tr>
<tr>
<td>Preferred</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table III: Significance levels from the analysis of variance of the experimental factors.

The subjects had a finite number of levels of dispersion to choose from (six). There was no indication on the computer display of the current level, except at the extremes, where the corresponding
button was visually and functionally disabled. The subjects may have been disinclined to choose a
dispersion level that they thus knew to be at an extreme.

Only unison choir sounds were used. It is possible that the tolerances and/or preferences would
be different in a polyphonic context.

CONCLUSION
Given the described implementations of pitch scatter and formant smear in syntheses of unison
choir sounds, the present experiment showed that
- pitch scatter was easier to assess than was formant smear,
- on the average, experienced listeners preferred pitch scatter to be between 0 and 5 cents
  and would tolerate scatter between 10 and 15 cents,
- on the average, listeners weakly preferred smear in formants F3-F5 to be between 2
  and 6%, and would tolerate smear up to 12%,
- the preferred dispersions depended weakly on vowel and voice category. The largest
  scatter and smear were preferred and tolerated on the vowel [u] in bass-like voices.

ACKNOWLEDGMENTS
The enthusiastic participation, both of the KTH choir singers, on whose voice properties our syn-
theses are based, and of the listener subjects, is gratefully acknowledged. I owe thanks to my col-
league Gunilla Carlsson for her work with the singing synthesizer and her assistance in making
formant frequency measurements. Anders Askenfelt and Johan Sundberg made valuable comments
on the manuscript. This work was done in the context of the project "Analysis and Synthesis of
Choir Singing", which is generously supported by the Bank of Sweden Tercentenary Foundation.

References
Carlsson, G. & Ternström, S. (forthcoming): "Formant frequencies of Swedish amateur choir singers in
speech and singing."
sis allowing expressive control," pp. 315-318 in Proc. of the Int. Computer Music Conference, Faculty of
Music, McGill University. ICMC Montreal.
Dolson, M. (1983): A Tracking Phase Vocoder and its Use in the Analysis of Ensemble Sounds, Ph.D. the-
sis, California Institute of Technology.
181-184.
69.
517-522.
STL-QPSR No. 2-3, pp. 93-105.
APPENDIX

Comments from the subjects

Choral standards. Even before the dummy run, most subjects spontaneously asked what sort of a choir they were supposed to be judging. After all, larger dispersion can be "tolerated" in a childrens' choir or in a senior citizens' choir than in an elite concert choir. They were asked to follow their personal taste without making allowances for choir type.

General comments. After completing the experiment, the subjects were asked whether they had any (unspecified) comments to make. All had found the experiment interesting and entertaining. Most subjects said it was easier to state preferences for scatter than for smearing. This is manifest in Fig. 2. Some subjects said they found it difficult to maintain same criterion throughout the test. This implies that training effects can be anticipated. However, since the order of presentation was randomized for each subject and then discarded, such effects are not traceable in the end result.

Synthesis quality. Several subjects volunteered that the fidelity of the syntheses was generally high. Several subjects noted that with excessive smear, the sounds lost their natural quality and sounded more synthetic. One subject thought the female syntheses sounded more like chest voice or boy voices than is normally encountered; another found the soprano voices to be the least natural. This was probably due not to inadequate formant data but rather to an inappropriate glottal source. Our current model of the glottal source, derived from speech synthesis, is static and does not work well at high levels of $F_o$. For the female voice categories, the glottal source was instead synthesized by addition of a $1/f^2$ series of harmonic partials, extending no higher than half the sampling rate, which was 16 kHz. The relative phases of the partials were adjusted individually to give a reasonable flow waveform. This strategy produces a glottal source with somewhat stronger high partials than is normal for female voices.