Solo to A Capella Conversion - Synthesizing Vocal Harmony from Lead Vocals

Paul Yaozhu Chan, Minghui Dong, SW Lee, Ling Cen, Haizhou Li

Department of Human Language Technology, Institute for Infocomm Research (I2R),
Agency for Science Technology and Research (A*STAR), Singapore

(ychan, mhdong, swylee, lcen, hli)@i2r.a-star.edu.sg

Abstract

This paper presents our work in the automatic synthesis of vocal harmony.

Nearly all existing innovations of the like to date either require some musical understanding/ability or allows for dissonances (i.e. non-harmonious or clashing intervals) at various parts of the song. This inspires us to develop a method that is able to automatically synthesize vocal harmony even for ordinary singers with a poor sense of rhythm.

We evaluated our method by means of spectrogram comparison and subjective listening tests. A spectrogram comparison of our method and two existing methods against that of the human voice shows that our method is least dissonant and most similar to natural human vocals. Subjective listening tests conducted separately for experts and non-experts in the field confirm that the vocal harmonies synthesized using our method sounds the best in terms of consonance, transition, as well as naturalness and appeal.

Index Terms: singing synthesis, vocal harmony, accompaniment, pitch interpretation, A Capella

1. Introduction

Harmony is a much generalized term in music, and it can thus refer to several things in musicology. The term vocal harmony often refers to melodic lines consonant to the lead vocals. This carries the accompaniment to the latter, which carries the main melody. This will be its definition throughout this paper, and it will be used interchangeably with the term accompaniment.

The correct addition of vocal harmony can significantly enhance the way an unaccompanied lead melody sounds. Furthermore, the exposed imperfections of an unaccompanied vocal lead may be transformed into pleasant sounding features when an accompaniment is added to it. One explanation of this, for example, is that imperfections in pitch introduce such dissonances, however, they are a special case of harmony, where all the overtones of both the notes are completely aligned, producing a similar effect to unison, and, thus, hardly achieving the effect of harmony.

An improvement to this method (method 458-II in table 1) corrects this problem to a certain degree with specification of the song key, and, with this information [3,4,5], allows for use of the major and minor 3rd intervals. However, even though clashes with the introduction of notes outside the natural key are resolved, similar clashes with notes within the key are not resolved.

Vocoders have been popular in music production since the 1970s, especially for the generation of robot-like vocals. [6] The Electro-Harmonix Voicebox [7] is one such vocoder and uses an instrumental (ideally, a guitar) input as the carrier and the human voice as a modulator to generate vocal harmony.[8]

In this arrangement (method AUX in table 1), the singer and instrumentalist (ideally, the same person) is tasked with the job of synchronization, eliminating the need for the machine to perform alignment. However, the harmony input requirements make this more applicable to trained musicians and unsuitable for singers without any special music ability.

![Image](https://via.placeholder.com/150)

Table 1 Comparison of current Automatic Harmony Synthesis methods against our proposed method

<table>
<thead>
<tr>
<th>Method</th>
<th>Accompaniment</th>
<th>Vocoder or Re-synthesized</th>
<th>Synchronization</th>
<th>Musical Ability or understanding</th>
<th>Other comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chord by MIDI Track mode</td>
<td>Guitar/Other</td>
<td>V</td>
<td>Manual</td>
<td>Pro</td>
<td>BLK</td>
</tr>
<tr>
<td>Kageyama's Karaoke Apparatus</td>
<td>Guitar/Other</td>
<td>V</td>
<td>Manual</td>
<td>Pro</td>
<td>BLK</td>
</tr>
<tr>
<td>Antare's Harmony Engine</td>
<td>Guitar/Other</td>
<td>V</td>
<td>Manual</td>
<td>Pro</td>
<td>BLK</td>
</tr>
<tr>
<td>Kano's Karaoke Player</td>
<td>Guitar/Other</td>
<td>V</td>
<td>Manual</td>
<td>Pro</td>
<td>BLK</td>
</tr>
<tr>
<td>Our Proposed method</td>
<td>Guitar/Other</td>
<td>V</td>
<td>Automatic</td>
<td>Pro</td>
<td>BLK</td>
</tr>
</tbody>
</table>

The current state-of-the-art methods, being Kageyama’s Karaoke Apparatus [9] and Antare’s Harmony Engine (in Chord by MIDI Track mode) [3], use more advanced re-synthesis techniques and is largely based on the same concept. This time, however, there is no input instrument, and the singer is required to synchronize with the backing track. [3] is more of a tool for song producers or sound engineers and synchronization is usually performed manually. [9] tailors to
subjects who do not have to be very musically inclined, but requires them to be able to have some sense of rhythm and be able to sing in time (synchronize) with the backing track.

In this work, we attempt to synthesize vocal harmony from lead vocals without the requirement of an auxiliary instrument or synchronization with a backing track, effectively achieving A Capella vocals from solo lead vocals. The harmony information is still required, but may be in the form of a MIDI file. Synchronisation is performed automatically using our reliable pitch synchronization method. This eliminates the need for manual synchronization or input of harmony information, making this more suitable for non-musicians, but introduces the need for synchronization or alignment.

2. Method

Figure 1 summarizes our proposed method. The interpretation of the pitch trace of the lead vocal input is aligned with the MIDI pitch trace, and the alignment information is used to re-align the MIDI interval trace, which is derived from the ratio of lead and accompaniment tracks. After re-alignment, the MIDI interval trace is now synchronised with the interpretation of the lead vocal input and the vectors may be multiplied to derive the target pitch traces for synthesizing the vocal accompaniments. These are fed into a high quality speech synthesizer, STRAIGHT [10], together with the the original lead vocal input, which is then analyzed and re-synthesized. The outputs of the synthesis stage are weighted differently and summed into two separate channels (figure 1 shows one for simplicity) to get our stereophonic harmonized vocals. The lead vocals do not need to be re-synthesized, but we do so in our experiments for consistency.

Figure 1 shows a simplified flowchart of our harmony synthesizer.

2.1. Pitch Interpretation

The successful implementation of this method is highly dependent on the accuracy of pitch interpretation.

It is very difficult to design a reliable pitch interpretation stage because of the incoherent relationships between human pitch production, computer pitch detection, the actual pitch in the medium and human pitch perception. Hence our experiments are a good test of the robustness of our pitch-interpretation algorithm.

In this work, interpretation is carried out as an independent stage. As such, none of the data in the MIDI file is used in the guesswork. The main reason for this is that this work is part of a larger project that aims to accomplish the similar task without the provision of the MIDI file.

2.1.1. Pitch Derivation

Pitch derivation is performed by means of autocorrelation. This stage also serves as the preliminary Voiced/Unvoiced (V/U) discriminator since segments with undefined pitch may be identified as unvoiced segments at this point. Other means of V/U discrimination have been found to be unnecessary with the effectiveness of the V/U Correction stage.

2.1.2. V/U Correction and Octave Correction

Voiced/Unvoiced correction is next performed to correct transients of unvoiced misinterpretations in voiced speech (VUV) and vice-versa (UVU). VUV errors have to be corrected before UVU ones to preserve the accuracy of the transition locations, and during which, the pitch data at the unvoiced transients have to be interpolated. Linear interpolation is found to be more effective than cubic-spline interpolation, which is commonly considered to be more natural. This stage should be performed before any octave (8ve) correction is carried out.

After this, 8ve-correction is then performed using a similar method to identify and correct any octave transients.

2.1.3. Translation to Logarithmic MIDI Note-Number Scale

Translation to the MIDI Note-Number Scale is then performed using the formula:

$$n_{\text{midi-scale}} = 9 + \left(12\log_2\left(\frac{f_H \times 32}{440}\right)\right) \quad (1)$$

Where $f_H$ is frequency in Hz.

Unlike the MIDI Note Numbers which are discrete, however, the translated pitch values are unrounded and, thus, left continuous.

2.1.4. Estimation of Overall Tuning Drift

Perfect pitch refers to the ability of a person to remember and identify or sing a pitch without the need of a reference pitch. This is an ability that comes to very few people and even amongst the most professional singers, few have this ability. Thus, there often is a significant discrepancy between the actual overall average tuning of a singer and the corresponding key especially when he or she is singing without a reference pitch.

The overall tuning drift is initially estimated by taking the ‘circular average’ of the decimal parts of the voiced pitch.

2.1.5. Key and Note Values Determination

The overall tuning drift is subtracted from each note value, and the result is rounded to establish the initial note values.
The frequency of occurrence is tabulated for each note (figure 2a), where octaves of the same note are considered to be the same note. Each note is weighted differently for each key and the weighted sum of all notes is established (figure 2b) for each of the 12 possible musical keys. In this way, the most probable song key is established.

2.1.6. Correction of Accidentals

Accidentals indicate if a note used is common in the key of the particular song. Occasionally, a song might use notes outside its native key, but this is rare for most commercial styles. At this stage, it is assumed that all notes keep within the key, and notes that were previously rounded to accidental notes are further rounded to the next nearest note within the key.

It is recommended that this stage be omitted for styles such as jazz, where accidentals are common. The key weightings used also have to be different for minor keys and blues.

2.1.7. Rule based Transient Segment Correction

At this point, the pitch trace is somewhat established, with the exception of several transient segments. These transient segments should not be disregarded because of their significant contribution to the slight mis-alignment that accounts for prominent distortions in the final synthesized vocals. They are usually intended to take the pitch of sustained segments just before or after the transient, with the split point defining the point of transition between notes. Occasionally, they may also intended to take the pitch of the sustained mean or median of the transient. In the case of the former, the precise interpretation of the point of singer-intended transition is important for the proper alignment and segmentation of the voice, and ultimately the quality of the synthesized vocal harmony.

The transient segments are first identified based on lengths. Extremely short spikes of usually one or two frame-lengths are identified and removed. Nodal cues are extracted from pitch and amplitude envelope gradients as well as pitch and amplitude envelope peaks.

Finally, rules are established by a human expert in the field of music systems engineering in a systemic 'node and determinant approach’. Determinants are drawn from geometric cues such as pitch boundary, the states of the trailing and preceding segments and the pitch, amplitude and temporal proximity of each point to each boundary. Rules are then established by mapping the state of the determinants to the established nodes. New determinants, nodal points and exceptions are allowed in overlapping intersections.

2.2. Alignment

The pitch trace for the lead melody is first plotted by referring to the notation information in the MIDI file. The pitch trace of the actual lead vocals is automatically transposed to match the key of that of the MIDI file. Each point on a pitch trace is compared with each point in the other in the plotting of an $L_{\text{sync}}$ by $L_{\text{mid}}$ Matrix, with each cell containing the difference between both pitches. A perfect match would hence be represented by 0, and the greater the distance the value is away from zero, the greater the mismatch. The matrix is traversed from point (0,0) to point $(L_{\text{sync}},L_{\text{mid}})$, over the lowest mismatching trajectory.

2.3. Re-Synthesis

Figure 3 describes our method of re-synthesis. STRAIGHT [10], originally a high quality synthesizer for speech, is, here, used in the re-synthesis of the accompaniment vocals. The lead vocal input is analysed and re-synthesized according to the synchronized pitch-trace obtained after the re-alignment stage.

3. Experiment & Discussion

The lead vocals of Brahms' Cradlesong and Twinkle Twinkle Little Star were recorded and a MIDI file containing the sequencing of their lead melody and accompaniment(s) were supplied.

Figure 4 shows the first stanza of the contents of the MIDI sequencing of the song Brahms' Cradlesong in the transcribed format of a music score. This song is sequenced in three-part harmony (1 lead + 2 accompaniments), while the second song is sequenced in two (1 lead + 1 accompaniment).

Figure 5 shows the pitch trace before (blue) and after (red) the interpretation stage.

3.1. The Interpretation Stage

Figure 5 shows the pitch trace before (blue) and after (red) the interpretation stage.

Their x-coordinate similarity is an approximate indication of the effectiveness of the interpretation algorithm.

Figure 6 compares the midi pitch trace with that of the interpretation of the sung vocal lead. Their y-coordinate similarity is an approximate indication of the effectiveness of the interpretation algorithm.
3.2. The Alignment Stage

The matrix in figure 7 shows an $L_{\text{sung}} \times L_{\text{ MIDI}}$ matrix for pitch trace alignment. The plot on the left in green represents the MIDI pitch trace while the one at the bottom in red represents the pitch trace of the sung vocal lead after being refined by the interpretation algorithm. In the matrix itself, the cells coloured green represent mismatches where points along the MIDI trace are greater than those along the actual vocals trace, with cells in duller green representing greater mismatch. Similarly, the red cells represent points where the pitch of the actual vocals is greater than that of the MIDI pilot. Black cells denote a complete mismatch, where there are unvoiced or silent segments along the actual vocals. Cyan cells denote a perfect match. The magenta line that traverses the matrix represents the optimum low-cost short-path trajectory, which is our alignment information.

3.3. The Re-synthesis Stage and Final Outputs

We compared our method (S2A) with two of the methods mentioned in Section 1. The 458 method uses transpositions a perfect 4\textsuperscript{th}, a perfect 5\textsuperscript{th} or an octave away from the lead vocals as the harmony line(s). The KTV method emulates the effect of a singer singing slightly off-timing into the karaoke harmony device mentioned [9]. The spectrograms of the results are compared against that of the human voice. Listening tests are carried out to compare the 3 results.
transposing the fundamental up a perfect 5th, ‘D2’ identifies regions of dissonance or potential dissonance due to key or chord ignorance. ‘D1’ identifies regions of dissonance or potential dissonance due to timing inaccuracies. Finally, ‘E’ indicates incorrect points of transition due to misalignment.

The green and yellow ‘+’s, and red ‘−’s compliment ‘D1’ and ‘D2’ by indicating regions of consonance, coincidental consonance, and dissonance respectively. Coincidental consonance refers to less common regions where the alignment is completely off but consonance is observed even though unplanned. At indications of dissonance, the red ‘−’ coincides with consonant locations.

It may also be observed that of the three, the S2A spectrogram is most identical to the human voice.

3.3.2. Subjective Listening Tests

The songs “Brahms’ Cradlesong” and “Twinkle Twinkle Little Star” were synthesized using the 3 methods. For the first song, 2 accompaniments are synthesized for the second song, 1 accompaniment is synthesized.

For synthesis using the 458 methods, a perfect 4th below and an octave above were chosen for the first song and a perfect 5th above was chosen for the 2nd song. For synthesis using the KTV method, results are expected to differ greatly depending on the timing drift of the singer and it is difficult to identify a singer with the generic sense of timing. As such, a singer slightly (up to about 0.3 sec) out of timing is emulated as an example. This is done by setting a loose alignment criteria.

The 6 songs are made available for evaluation at http://www1.i2r.a-star.edu.sg/~mhdong/survey/.

3.3.2.1 By Vocal Experts

In the first test, 11 vocal experts were tasked to listen to the 6 songs and evaluate them in terms of consonance (harmony) and smoothness of transition.

These two characteristics were explicitly specified because of the following reasons:

- The 458 method, deriving the accompaniment by transposing the lead vocals by a fixed interval throughout the song, is expected to score well in terms of smoothness of transition but suffer in terms of consonance.

- The KTV method, on the other hand, deriving its accompaniment from midi whilst relying on manual synchronization, is expected to score better in terms of consonance but poorly in terms of transition. However, it is anticipated that poor location of transition can have a negative affect on its score in consonance.

Table 2 shows their average ratings for each of the songs on a scale of 1 to 5.

<table>
<thead>
<tr>
<th>Consonance / Harmony</th>
<th>458</th>
<th>KTV</th>
<th>S2A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brahms’ Cradlesong</td>
<td>2.8</td>
<td>1.5</td>
<td>4.4</td>
</tr>
<tr>
<td>Twinkle Twinkle Little Star</td>
<td>3.8</td>
<td>1.8</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>3.5/5.0</td>
<td>1.6/5.0</td>
<td>4.5/5.0</td>
</tr>
<tr>
<td></td>
<td>(66.4%)</td>
<td>(32.7%)</td>
<td>(90.0%)</td>
</tr>
</tbody>
</table>

As anticipated, the S2A method performs best in terms of both consonance and smoothness of transition. This result verifies the effectiveness of our method.

We initially expected the transition score for the 458 method to be higher than our method. We expect that this result may be attributed to the unnatural sounding effect at the transitions when each part transits in perfect synchrony.

We also expect that the consonance score for the 458 method might be slightly biased towards the positive direction because certain clashes/dissonance might not be obvious in the absence of a backing track.

3.3.2.2 By Non-Experts

In the second test 12 non-experts were tasked to listen to the 6 songs. Because non-experts are not expected to be as attentive to aural detail, we tasked them to rate each song on a scale of 1 to 10 on how pleasant and natural they thought each song sounded. Table 3 lists their ratings on a scale of 1 to 10.

<table>
<thead>
<tr>
<th>Pleasant / Natural Sounding</th>
<th>458</th>
<th>KTV</th>
<th>S2A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brahms’ Cradlesong</td>
<td>6.7</td>
<td>6</td>
<td>8.25</td>
</tr>
<tr>
<td>Twinkle Twinkle Little Star</td>
<td>5.5</td>
<td>4.7</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>6.1/10</td>
<td>5.3/10</td>
<td>7.0/10</td>
</tr>
<tr>
<td></td>
<td>(60.8%)</td>
<td>(53.3%)</td>
<td>(70.0%)</td>
</tr>
</tbody>
</table>

The results this time, once again verifies the effectiveness of our method, although the results are not as obvious due to the lack of attention to aural detail of the non-experts.

4. Future Work

This is part of a greater piece of work that includes a rhythm interpretation stage as well as a harmony calculator stage that will eliminate the needs for the alignment and the MIDI prerequisite respectively. With the success of the melody interpretation in this project, we will move on to develop a versatile rhythm stage in the next phase of our research.

5. Conclusions

In this paper, we have proposed a new method of automatic vocal harmony that is significant because, unlike existing methods, it is suitable for singers without a good sense of rhythm yet does not sacrifice the quality of consonance. Spectrograms as well as subjective listening tests by field experts and non-experts indicate the successfullness of our method in achieving a better level of perceived harmonic consonance, transitional smoothness, as well as overall naturalness and pleasantness.
6. References


