SOLO TO A CAPELLA CONVERSION
- SYNTHESIZING VOCAL HARMONY FROM LEAD VOCALS

Paul Y Chan, Minghui Dong, S W Lee, Ling Cen{ychan, mhdong, swylee, lcen}@i2r.a-star.edu.sg

ABSTRACT

This paper presents our work in the automatic synthesis of vocal harmony. Existing innovations either allow for dissonances (i.e. non-harmonious or clashing intervals) at various locations or require some musical ability of the user. We have developed a method that is able to automatically synthesize vocal harmony even for ordinary singers with a poor sense of harmony and rhythm. We have evaluated our method by means of spectrogram comparison as well as subjective listening tests. A spectrogram comparison of our method and two popular 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 harmony synthesized using our method sounds the best in terms of consonance, inter-syllable transition, as well as naturalness and appeal.

Index Terms— Singing synthesis, vocal harmony, accompaniment, pitch interpretation, pitch alignment, A Capella

1. INTRODUCTION

The term vocal harmony often refers to melodic lines that are to be sung 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 illustration of this, for example, is the way harmonic phase discrepancies between lead and accompaniment vocals translate into perceived amplitude and frequency variations which are perceived to be interesting to the human ear. This is one of the reasons why vocal harmony is so popular in the production of commercial music.

However, unlike lead melodies, vocal accompaniment melodies are often difficult for most people to learn. It is not uncommon even for professional singers to have to spend time rehearsing beforehand. This inspired a variety of vocal harmony synthesis methods.

Believed to be originating from the Gregorian Chants [1,2], a traditional method (referred to as 458 in table 1) for the derivation of the harmony line (accompaniment melody) from the lead vocal line is by indiscriminately using perfect 4th or 5th or octave (8ve) intervals [3,4,5]. Perfect 4th and 5th intervals, however, in contemporary music, introduces potential dissonances with the 4th and 7th notes in the most common major scale, which are undesirably sharpened and flattened, respectively. In the minor scale, they can introduce a variety of dissonances, depending on the type of minor scale. Octave intervals do not introduce such dissonances since they are a special case of harmony, where all the overtones of both the notes are completely aligned. However, this produces an effect very similar to perfect unison, which hardly achieves the effect of harmony.

An improvement to this method (referred to as 458-II in table 1) partially corrects this problem with the requirement of user specification of the song key. This information [3,4,5], allows for use of the major and minor 3rd intervals. However, even as clashes with the introduction of notes outside the natural key are resolved, clashes with notes within the key cannot be resolved this way.

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.

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. Moreover, since [3] is more of a tool for song producers or sound engineers, synchronization usually requires manual correction after recording. [9] sets forth to tailor to subjects who do not have to be very musically inclined, but requires them to be able to have some sense of rhythm, that is, to be able to sing in time (manually 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. Synchronization 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 relationship between the lead and accompaniment tracks. After re-alignment, the MIDI interval trace is now synchronized 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 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.

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 or any harmonizing data.

2.1.1. Pitch Derivation

Primary pitch derivation is performed by means of autocorrelation. [11] This stage also serves as the preliminary Voiced/Unvoiced (V/U) discriminator [12] since segments with undefined pitch may be identified as unvoiced segments at this point. Other means of V/U discrimination has been found to be unnecessary with the effectiveness of the V/U Correction stage.

MIDI interval trace, which is derived from the relationship between the lead and accompaniment tracks. After re-alignment, the MIDI interval trace is now synchronized 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.

2.1.2. V/U Correction and Octave Correction

Voiced/Unvoiced correction [12] 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. 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 (8\text{th}) correction is carried out. Octave correction is then performed using a similar method to identify and correct any octave transients.

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

<table>
<thead>
<tr>
<th>Device/Method</th>
<th>How existing methods compare with our proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Aux 458 458-II KTV S2A</td>
</tr>
<tr>
<td>Accompaniment Derivation</td>
<td>Guitar/Other KB Fixed interval from lead vocal Fixed interval from lead with exceptions MIDI MIDI</td>
</tr>
<tr>
<td>Vocoded or Re-synthesized</td>
<td>Usually Vocoded or Pitch-Shifted except HE1–3 Re-Synthesized</td>
</tr>
<tr>
<td>Synchronization</td>
<td>Manual Not Applicable Not Applicable Manual Auto</td>
</tr>
<tr>
<td>Dissonance / ‘wrong’ notes</td>
<td>Min Common, Key clashes Common, Chord clashes None Min Almost none</td>
</tr>
<tr>
<td>Musical Ability or Understanding</td>
<td>Guitar / Keyboard Pro Pro Min Pro Min Pro Synch None</td>
</tr>
</tbody>
</table>

Other comments

Algorithm:
Aux: Auxiliary input of harmony information
458: Blind fixed-interval[1] (usually 4th, 5th or 8\text{th}) apart from lead vocal
458-II: 458 but avoids Type1 clashes & allows 3rd intervals
KTV: Harmony from score/MIDI
S2A: Harmony from score/MIDI with automatic alignment

Figure 1 A simplified flowchart of our harmony synthesizer

2.1.2. V/U Correction and Octave Correction

Voiced/Unvoiced correction [12] 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. 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 (8\text{th}) correction is carried out. Octave 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:

\[
 n_{\text{midiscale}} = 9 + 12\log_2 \left( \frac{f_{\text{Hz}} \times 32}{440} \right)
\]  
(1)

where \( f_{\text{Hz}} \) is frequency in Hz. [13]

Unlike the MIDI Note Numbers which are discrete, however, the translated pitch values are unrounded and 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, 14]

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 may be inconsistent. The key weightings used may have to be modified for other scales such as minor and blues.

2.1.7. Rule-Based Transient Segment Correction

At this point, the pitch trace is almost established, with the exception of several transient segments. These transient segments should not be disregarded because of their contribution to misalignments that account for distortions in the final synthesized vocals. While they are more often 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, they may occasionally also be 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 nodal points (exceptions), and corresponding determinants, 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 \( L_{\text{Sung}} \) by \( L_{\text{midis}} \) 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. For each point in the matrix, individual lowest its cost distance away from the destination point \( (L_{\text{Sung}}, L_{\text{midis}}) \) is then computed, with cost being the degree of match. The matrix is traversed from point \((0,0)\) to point \((L_{\text{Sung}}, L_{\text{midis}})\), choosing the lowest cost adjacent point at each step in a fashion very similar to [15]. This is further illustrated in section 3.2.

2.3. Re-Synthesis

Figure 3 describes our method of re-synthesis. We use STRAIGHT [10], originally a high quality speech synthesizer, here, in the re-synthesis of the singing voice. The lead vocal input is analyzed 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 sequence of the song Brahms' Cradlesong in the transcribed format of a music score. This arrangement of the song is for three-part harmony (1 lead + 2 accompaniments) while the arrangement used for the second song (not shown) is sequenced for two voices (1 lead + 1 accompaniment).

3.1. The Interpretation Stage

Figure 5 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.

Figure 6 shows the pitch trace raw (thin dark) and after (thick pale) the interpretation stage.

3.2. The Alignment Stage

The matrix in figure 7 shows an L_{sung} by L_{midi} matrix for pitch trace alignment. The plot on the left represents the MIDI pitch trace while the one at the bottom represents the pitch trace of the sung vocal lead after being refined by the interpretation algorithm. In the matrix itself, the darkness of the cells represent the match between points along the MIDI trace and those along the actual vocals trace. White cells denote a complete mismatch, or where there are unvoiced or silent segments along the actual vocals. The black 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 experiment uses transpositions a perfect 4th, a perfect 5th or an 8ve away from the lead vocals as the harmony line(s). The KTV experiment 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.

3.3.1. Spectrogram Plots

Figure 8 compares the spectrogram plots for the harmonization of the song “Twinkle Twinkle Little Star” using the 3 methods against that of the human voice. The last stanza is compared here, “How I wonder what you are”.

In the spectrograms, ‘A’ identifies the lead line and ‘B’ identifies the accompaniment line. ‘C’ cites an example of the undesirable effect of “perfect harmonic alignment” with the 458 method. It is undesirable for, as explained in 1, perfect phase alignment does not produce perceived frequency or amplitude variations which are musically appealing. [2] - here, the 3rd harmonic of the lead aligns almost perfectly with the 2nd harmonic of the accompaniment when the accompaniment is derived by transposing the fundamental up a perfect 5th. ‘D1’ identifies
regions of dissonance or potential dissonance due to key or chord ignorance. ‘D2’ identifies regions of dissonance or potential dissonance due to timing accuracies. Finally, ‘E’ indicates incorrect points of transition due to misalignment.

The ‘+’s, and ‘-’s compliment ‘D1’ and ‘D2’ by indicating regions of consonance or 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 ‘-’s coincide 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 loose alignment criteria.

The files for “Brahms’ Cradlesong” synthesized using all 3 methods are both submitted for evaluation together with this paper.

#### A. Vocal Experts’ Opinion

**Table 2** Results of Subjective Listening Tests by Vocal Experts for

<table>
<thead>
<tr>
<th></th>
<th>458</th>
<th>KTV</th>
<th>S2A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consonance / Harmony</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<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><strong>Smoothness of Transition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brahms’ Cradlesong</td>
<td>2.4</td>
<td>1.4</td>
<td>2.8</td>
</tr>
<tr>
<td>Twinkle Twinkle Little Star</td>
<td>2.5</td>
<td>1.8</td>
<td>3.4</td>
</tr>
</tbody>
</table>

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:

![Figure 8](image-url) Spectrograms of vocal harmony synthesized using (a) the 458 method, (b) the KTV method and (c) the S2A method (our method) against that of (d) actual human vocals.
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.

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 did not expect to outperform the 458 method to be higher than our method in terms of transitional smoothness. We expect that this result may be attributed to the unnatural effect produced by the 458 method’s synchronized transitions.

B. Non-Experts’ Opinion

In the second test12 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 3 Results of Subjective Listening Tests by Non-Experts

<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6.1/10</th>
<th>5.3/10</th>
<th>7.0/10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(60.8%)</td>
<td>(53.3%)</td>
<td>(70.0%)</td>
</tr>
</tbody>
</table>

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

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

4. 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 successfulness of our method in achieving a better level of perceived harmonic consonance, transitional smoothness, as well as overall naturalness and pleasantness.

5. 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.

6. REFERENCES