VOCADITO: A DATASET OF SOLO VOCALS WITH $F_0$, NOTE, AND LYRIC ANNOTATIONS – EXTENDED ABSTRACT

Rachel M. Bittner, Katherine Pasalo, Juan José Bosch, Gabriel Meseguer-Brocal, David Rubinstein

Abstract

To complement the existing set of datasets, we present a small dataset entitled vocadito, consisting of 40 short excerpts of monophonic singing, sung in 7 different languages by singers with varying levels of training, and recorded on a variety of devices. We provide several types of annotations, including $F_0$, lyrics, and two different note annotations. All annotations were created by musicians. In this extended abstract, we omit all analysis, and refer the reader to the extended technical report [1]. Vocadito is made freely available for public use.

1. INTRODUCTION

The singing voice is one of the most expressive instruments, and one that is particularly challenging to transcribe [2]. A common task is to transcribe a voice’s pitch content, either in the form of frame-level $f_0$, or in the form of note-events. These two representations are related, but not trivial to convert between. Given a recording’s frame-level $f_0$, one cannot trivially create note events by e.g. quantizing because it lacks information about onsets, and it is ambiguous how to group pitches into events. Similarly, it is not possible to infer what the frame-level $f_0$ is for a recording given a sequence of notes, as expressive performance information such as vibrato or glissando is not encoded.

Notes themselves are known to have a degree of subjectivity - (as we’ll see further results for in this paper) - given the same recording, two humans may not generate the same sequence of note events. For this reason, when evaluating the correctness of estimated note events, it is common to allow a tolerance window for where an onset is placed, and an ever larger tolerance for where an offset is placed. For other related tasks, e.g. chord recognition [3] and music segmentation [4], datasets with multiple annotations have been created to address the inherent subjectivity of the task. To date, no such dataset exists for note annotations.

Few datasets exist which contain both human-annotated note and $f_0$ data, making it difficult to study interactions between them. The same is true for lyrics and note or $f_0$ data. Table 1 provides an overview of existing datasets with solo or monophonic singing voice. While there are a number of existing datasets with note, $f_0$ or lyric annotations, when it comes to any one task there are actually only a few. When it comes to note estimation for note estimation for solo singing voice, only 3 exist: Molina, DALI_multi, and TONAS. DALI_multi is large, however the annotations are crowdsourced and automatically aligned – while this is useful for training, it is not an appropriate dataset for evaluation. Molina and TONAS are both good evaluations sets, but (like vocadito) are relatively small. We are further restricted if both $f_0$ and note annotations are needed - leaving only TONAS. Furthermore, no dataset provides more than one note annotation, and as we will see, the note annotation task itself is quite subjective.

In this extended abstract, we describe the creation of the vocadito dataset. Vocadito is made freely available on Zenodo under a Creative Commons license, and is included in the mirdata library. For more details and analysis, please see the complete technical report [1].

2. DATASET CREATION

2.1 Data Collection

Audio recordings for vocadito were collected from 28 volunteers, with varying singing experience. In order to simulate a “real-world” setting, we did not restrict volunteers to record using high-quality microphones, and many of the recordings are from cell phone or computer microphones. Volunteers were asked to choose an original or public domain song (e.g. folk or children’s music), and create up to three 10-40 s recordings. Volunteers agreed to their recordings being anonymously included in this dataset and publicly released. We ensured that no composition is repeated across the 40 recordings. The collected recordings were manually edited to remove any long silences at the beginnings or ends, and reformatted from their original format into 44.1 kHz, 16 bit mono.wav files.

2.2 Human Annotations

We created four types of human-labeled annotations for Vocadito: frame-level $f_0$, notes, lyrics, and track-level metadata (e.g. the sung language). All annotators are experienced musicians.
Table 1. A (non-exhaustive) overview of existing datasets for solo vocals. **Polyphony** indicates how many voices are present at one time. A + indicates that there are recordings where multiple singers are singing at once. **Isolated** indicates whether the vocal recordings are isolated or not (there is background music).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Saraga Carnatic  [5]</td>
<td>1</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>249</td>
</tr>
<tr>
<td>DALI [6, 7]</td>
<td>1+</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>7756</td>
</tr>
<tr>
<td>cante100 [8, 9]</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>MIR-1K</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>1000</td>
</tr>
<tr>
<td>iKala [10]</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>252</td>
</tr>
<tr>
<td>MedleyDB [11, 12]</td>
<td>1</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>Jingju A Cappella [13]</td>
<td>1</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>82</td>
</tr>
<tr>
<td>VocalSet [14]</td>
<td>1</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>3560</td>
</tr>
<tr>
<td>ChoirSet [15]</td>
<td>4+</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>DALI_multi [16]</td>
<td>1+</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>513</td>
</tr>
<tr>
<td>Molina [17]</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>TONAS [18, 19]</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓ ✓ ✓</td>
<td></td>
<td>72</td>
</tr>
<tr>
<td><strong>vocadito</strong></td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓ ✓ ✓</td>
<td></td>
<td>40</td>
</tr>
</tbody>
</table>

Figure 1. Annotations for the first 5.5 seconds of track 6 of vocadito. The plot shows $f_0$ annotations (black), note annotations by annotators 1 and 2 in red/blue respectively (overlaps shown in purple), and lyrics above the plot. Lyric time-alignments were labeled here for demonstration purposes, but are not part of vocadito.

Frame-level $f_0$. $f_0$ annotations are created using Tony [21], a software application which first automatically estimates the $f_0$ using the pYIN [22] algorithm, and then allows an annotator to manually correct mistakes made by the algorithm. One annotator created $f_0$ annotations in this manner for each of the 40 tracks. The annotator reported that the majority of the corrections involved either removing $f_0$ estimates in frames where no $f_0$ was present (e.g. during consonants), or in adding missing $f_0$ estimates for frames with low pitch.

Notes. Note annotations were also created using Tony, which similarly for $f_0$, uses an algorithm to estimate note events and allows an annotator to correct the estimates. In order to explore the subjectivity of creating note events for vocals, two different annotators created separate note event annotations for each of the 40 tracks. The annotators were instructed to annotate the notes they would play if they had to reproduce it on the piano. Note that while the piano is restricted to a semitone grid, the annotation software (Tony) allows notes to have any continuous pitch; which is convenient, since singers did not necessarily sing in standard tuning (440 Hz). Thus, the instructions provided to the annotators regarding the piano refers more to how to segment notes in time than to how to label the pitch.

Lyrics. Lyric annotations were created by fluent speakers of the sung language for each song. Annotators listened to the recording and transcribed the words as they are sung exactly (even when this deviates slightly from the text of the original composition). Line breaks in the lyrics indicate the end of a musical phrase, and a blank line indicates the end of a musical section. The lyrics are provided as text, without timing information. All lyrics are written in the Latin alphabet. For the two tracks which are in Chinese, lyrics are provided in both Chinese characters and in pinyin. For the two tracks where more than one language is present, we indicate the language as language1+language2.

Metadata. We provide track-level metadata for each of the tracks, including (1) the sung language (2) the singer ID (anonymized) (3) the average pitch (computed from the $f_0$ annotations).
3. REFERENCES


