# Phoneme-Informed Note Segmentation of Monophonic Vocal Music

**Yukun Li** * **Emir Demirel** † **Polina Proutskova** **Simon Dixon**
Centre for Digital Music,
Queen Mary University of London, UK
yukun.li@qmul.ac.uk

## Abstract

Note segmentation of vocal pitch tracks is an inherently difficult problem, on which human judgments often disagree. We propose a novel note segmentation method that leverages phonemic information. Phonemes and pitch tracks are automatically extracted and jointly utilised to estimate note transition regions. Note onsets are determined within these regions using an onset detection function. Finally, an HMM-based note tracker adds further note boundaries for the case where multiple notes are sung on the same vowel. Our note segmentation method outperforms the previous best method on a standard public test set, and is shown to be somewhat robust against different types of lyrical content. Because its performance is less convincing on another dataset, we analyse problem cases and suggest possible confounding issues.

## 1 Introduction

Automatic music transcription refers to converting an acoustic waveform into a symbolic representation. While monophonic instrument transcription is often considered to be a solved problem in music information retrieval (Benetos et al., 2013), this is not the case for singing, where pitch is rarely stable (Dai and Dixon, 2019).

A singing transcription system usually consists of two main steps: pitch tracking and note segmentation. Firstly, the pitch and voicing are estimated at each time point in the audio; secondly, the continuous pitch track is segmented into notes which have onset, offset and an indicative pitch. For the first step, we use the PYIN algorithm (Mauch and Dixon, 2014), which improves on the widely used YIN algorithm (de Cheveigné and Kawahara, 2002) for estimating the fundamental frequency and voicing (presence or absence of pitch) of a monophonic signal. As PYIN works well for monophonic pitch estimation, we only focus on note segmentation in this paper.

Despite the high level of research activity in this area, the average F-measures of note-level transcription metrics (Correct Onset, Pitch and Offset, COnPOff (Molina et al., 2014)) obtained by state-of-the-art systems are all lower than 60%. Detection of "soft" onsets and offsets is still an unsolved problem in note segmentation. Soft onsets and offsets occur when adjacent notes are smoothly connected without obvious loudness variations. In most cases, however, there is a phonetic change between notes. Various spectral features have been used to detect timbre changes, either by selecting as boundaries peaks above a threshold in the measure of timbre change (Gómez and Bonada, 2013; Yang et al., 2017), or by modelling vowels and their transitions using an HMM (Hsuan-Huei Shih et al., 2002; Heo and Lee, 2017). More recently, utilising the flexibility of deep neural networks, Fu and Su (2019) augmented their input data with onset- and offset-related features to improve note segmentation and transcription performance.

To solve the problem of soft onsets and offsets, this paper investigates whether phonemes extracted by a state-of-the-art automatic lyrics transcription system (Demirel et al., 2020) can make a positive contribution. We hypothesise that phoneme information can be used to narrow down the range of frames where onsets and offsets are likely to occur. In particular, consonants are possible indicators of note boundaries, whereas vowels, unless there is a significant change of pitch or of the vowel, indicate the body of a note.

## 2 Method

Based on the annotation approach of Molina et al. (2014), our method assumes that note boundaries

*YL is supported by a China Scholarship Council and Queen Mary University of London joint Ph.D. Scholarship.

†ED receives funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

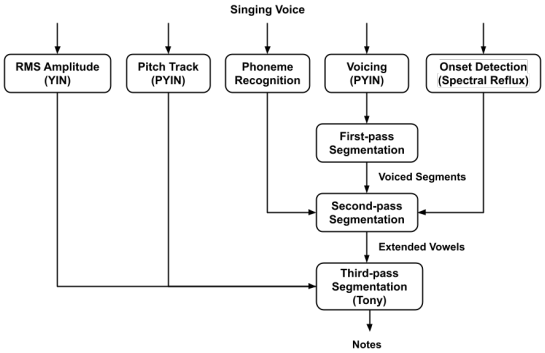
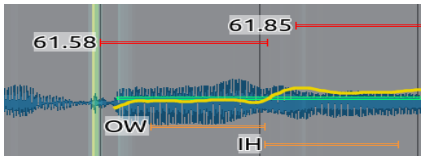
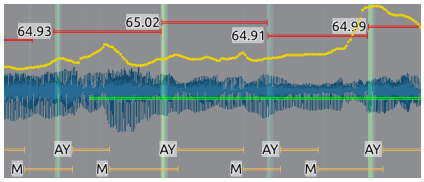



Figure 1: Proposed 3-step note segmentation method.




(a) Adjacent vowels with similar pitches erroneously merged into a single note.




(b) Successive notes sung on similar pitches with voiced phonemes between vowels, leading to multiple merge errors.

Figure 2: Examples of common errors made by the Tony software (Mauch et al., 2015). The ground truth segmentation (red) is labelled with median pitch in semitones (MIDI).[3] The PYIN pitch track is yellow, the note region extracted by Tony is bright green, detected phoneme boundaries are orange, and spectral flux is represented by the brightness of vertical lines.

can be categorised into three types: (1) the beginnings and ends of voiced segments; (2) phonetic changes; (3) pitch[1] and amplitude changes. We detect these types of note boundaries and segment the vocal track in a three-step cascading approach which produces successively finer segmentations at each step (Figure 1).

In Step 1, voiced segments (segments of continuous pitch activity) are determined, based on the PYIN pitch track. In Step 2, the voiced segments are further segmented based on phonetic change, to create what we call *extended vowel* regions, as described in Section 2.1. In Step 3, extended vowel segments are further divided based on pitch and amplitude changes given by the PYIN algorithm. The main novelty of this approach is the incorporation of phonetic information into an existing framework for note segmentation, through the introduction of the second step, which we now describe in detail. [2]

### 2.1 Step 2: Phoneme-Informed Segmentation

In order to detect phonetic change, the phonemes are automatically transcribed and temporally aligned using a state of the art lyrics transcription system (Demirel et al., 2020). The Spectral Reflux onset detection function (Sapp, 2006) is then used to estimate the note boundaries more precisely.

Demirel et al.'s system provides a transcribed phoneme sequence with aligned timings, but it claims a boundary accuracy tolerance of 50 ms. To detect note boundaries more precisely, we fine-tune the phonetic output with a simple additional signal processing step. First, we categorise the phonemes into vowels and consonants, determining the inter-vowel regions. We expand the inter-vowel regions by 50 ms each side to account for the system's tolerance. Finally, the maximum of spectral reflux in the expanded inter-vowel region determines the exact note boundary. The pitch of each segment is calculated as the median of the pitch track within the segment.

Figure 2 illustrates the need for this step, showing examples where Tony, a benchmark method for monophonic singing voice transcription (Mauch et al., 2015), makes the systematic error of under-segmentation of successive notes having continuous steady pitch tracks during note transitions. These instances occur when consecutive notes are sung either with no consonants or silent gaps (breathing, articulation, etc.), or with short voiced consonants, between the successive vowels. When there are two adjacent vowels with no gap in between (Figure 2a), the note boundary is determined by the timing of the vowel transition. Where there is a gap between consecutive vowels (Figure 2b), we determine the note boundary as the location of the local maximum of the spectral reflux between the vowels in question.

### 2.2 Step 3: Pitch and Amplitude Changes

Steps 1 and 2 detect inter-vowel note boundaries, but there are also note boundaries within vowels that are communicated via pitch and amplitude changes. In such cases, phoneme-based segmen-

[1] We follow PYIN (Mauch et al., 2015) in setting the threshold of pitch change required for a note boundary to $\frac{2}{3}$ of a semitone.

[2] Step 1 is simple and does not require further description.

[3] We use non-integer values to represent continuous pitch.

tation is unable to determine the note boundaries. To estimate the timings of such boundaries, we apply the HMM-based segmentation method of Tony (Mauch et al., 2015) within the extended vowel regions resulting from steps 1 and 2 (Fig. 2).

## 3 Evaluation

For evaluation, we use the framework proposed by Molina et al. (2014), including their monophonic singing dataset (38 recordings with total duration 1154 seconds). We first show results from an ablation study using standard metrics on this dataset, and then follow this up with a comparison with recently published systems. We then examine the effect of linguistic properties of the data on segmentation results, and test on another publicly available dataset (Dai et al., 2015).

### 3.1 Ablation Experiments

To illustrate the contribution of each step in our approach to the overall performance, we report results for different combinations of the steps described in Section 2. Table 1 lists the methods, the features used, and their performance on three metrics. Since we consider voicing analysis (Step 1) as fundamental to any singing segmentation approach, we always include this feature, and test different combinations of Steps 2 and 3.

The first three evaluation metrics (columns) in Table 1 are the F-measures of COnPOff, COnP and COn, as used in MIREX, and the other three are the count proportions for various types of segmentation errors (Molina et al., 2014). COnPOff measures the rate of transcribed notes with correct onset ($\pm$50 ms), pitch ($\pm$0.5 semitones) and offset ($\pm$50 ms or $\pm$20% of the duration of the reference note). COnP represents correct onset and pitch, and COn evaluates correct onset only. A "Split" error means the ground truth note is split into multiple notes in the transcription, while a "Merged" error is the opposite. A "Spurious" note error occurs when a transcribed note does not overlap in time with any ground truth note. The results indicate that the various components of our approach each contribute positively to the overall performance on all three note-level metrics. In particular, disabling either Step 2 or 3 reduces performance by $\sim$9% on the strictest measure, with Step 2 making the greater contribution to the results.

In addition, for all versions of the system, relaxing the requirement of correct offset detection results in 15–20% better results, whereas relaxing the requirement to estimate the correct pitch only contributes a further 5% to the results. The remaining errors (relating to the onset) account for 20–25% of the results, so it is clear that a high proportion of errors relate to onsets and offsets, or in other words, the segmentation.

### 3.2 Comparison to the State-of-the-Art

In Table 2 we compare our results to published work on singing transcription. We tested our three-step method on Molina et al.'s dataset[4] (Molina et al., 2014), and compared its performance with six of the previous best sung note segmentation and transcription methods.

Overall, the results demonstrate that our proposed method achieves the best overall performance (F-measure), by a small margin over Fu and Su's recent work (Fu and Su, 2019). In addition, we have the lowest rates of merged and spurious note errors, and only on the split error metric are our results inferior to other systems. This means that the system has a tendency to over-segment the sung notes, compared to other published work. Looking more closely at the results, however, we see that most of the systems with lower split errors have very high rates of merged errors, so they are in fact under-segmenting the signal.

### 3.3 The Effect of Language

In this subsection, we investigate the robustness of the proposed system to various types of lyric content, including different languages and non-linguistic content. In addition, we discuss the sources of errors made by our system. We categorised the dataset by Molina et al. (2014) into the five groups represented by the columns of Table 3. Melodies in this dataset are sung either in English, Spanish and/or the following isolated syllables: /Na/, /Da/ and /La/. Using the F-measure of COnPOff, we compare performance of three versions of our system.

Several surprising results appear in Table 3. Starting with the complete system (the final row), the results for Spanish are about 19% higher than those for English. It is not entirely unexpected that Spanish is easier to segment, but this should be weighed against the fact that the phoneme predictions come from a lyrics transcriber that is trained on English language songs (Demirel et al., 2020).

[4]For methodological correctness, we exclude 3 samples during evaluation which had been used during analysis and development, even though we did not tune any hyperparameters on these samples. The results do not change substantially between the two versions of the dataset.

| Methods | Features Used | COnPOff | COnP | COn | Split | Merged | Spurious |
|---|---|---|---|---|---|---|---|
| Steps 1+2 | voicing(1), phoneme(2), onset(2) | 0.525 | 0.712 | 0.761 | **0.013** | 0.235 | 0.128 |
| Steps 1+3 | voicing(1), pitch(3), amplitude(3) | 0.520 | 0.683 | 0.741 | 0.079 | 0.233 | 0.114 |
| Steps 1+2+3 | voicing(1), phoneme(2), onset(2), pitch(3), amplitude(3) | **0.610** | **0.762** | **0.807** | 0.093 | **0.078** | **0.035** |

Table 1: Transcription performance (F-measure) on the Molina dataset for three versions of our approach. Columns represent correct (C) onset (On), pitch (P) and/or offset (Off), respectively, and three types of error (see Sec. 3.1).

| Method | Precision | Recall | F-measure | Split | Merged | Spurious |
|---|---|---|---|---|---|---|
| Ryynänen & Klapuri (Ryynänen and Klapuri, 2004) | 0.304 | 0.315 | 0.308 | 0.105 | 0.248 | 0.116 |
| Gómez & Bonada (Gómez and Bonada, 2013) | 0.430 | 0.373 | 0.398 | 0.140 | 0.167 | 0.071 |
| Molina et al. (SiPTH) (Molina et al., 2015) | 0.397 | 0.440 | 0.415 | 0.074 | 0.309 | 0.157 |
| Yang et al. (Yang et al., 2017) | 0.409 | 0.436 | 0.421 | 0.064 | 0.230 | 0.120 |
| Mauch et al. (Tony) (Mauch et al., 2015) | 0.510 | 0.534 | 0.520 | 0.079 | 0.230 | 0.112 |
| Fu and Su (Fu and Su, 2019) | 0.625 | 0.569 | 0.594 | **0.048** | 0.080 | 0.044 |
| Steps 1+2+3 (whole dataset) | **0.626** | **0.597** | **0.610** | 0.093 | **0.078** | **0.035** |
| Steps 1+2+3 (test set) | 0.634 | 0.606 | 0.618 | 0.090 | 0.080 | 0.035 |

Table 2: Transcription and segmentation performance on the whole dataset of Molina et al. (2014), compared with published results (best results in bold). The first three rows are reported by Molina et al. (2015), and the following two are quoted from Yang et al. (2017). The final row compares performance evaluated on the smaller test set. The first three columns refer to COnPOff (correct pitch, onset and offset) results; the remaining columns are segmentation error types (see Sec. 3.1).

| | English | Spanish | /Na/ and /La/ | /Da/ and /La/ | Syllable and Lyrics Mixed |
|---|---|---|---|---|---|
| Number of recordings | 10 | 15 | 7 | 1 | 5 |
| Steps 1+2 | **0.612** | 0.609 | 0.325 | 0.178 | 0.448 |
| Steps 1+3 | 0.443 | 0.602 | 0.396 | 0.652 | 0.575 |
| Steps 1+2+3 | 0.523 | **0.709** | **0.520** | **0.677** | **0.596** |

Table 3: Comparison of transcription performance (F-measure of COnPOff) for different categories of lyrics.

| Methods | COnPOff (F-measure) | Merged | Split |
|---|---|---|---|
| Step 1 | **0.645** | 0.023 | **0.004** |
| Steps 1+2 | 0.603 | 0.018 | 0.045 |
| Steps 1+2+3 | 0.614 | **0.005** | 0.069 |

Table 4: Transcription and segmentation performance comparison for the dataset of Dai et al. (2015).

Since the English language songs match the training data quite well, there are very few merge errors in these songs after Step 2, and the subsequent segmentation causes over-segmentation and degrades performance. In other cases, Step 3 improves performance, especially for non-linguistic samples.

For non-linguistic content, we are wary of making strong claims as the amount of data is quite small. We observe that the Step 2 output is considerably worse with non-linguistic syllables than on songs with linguistic content, but this difference is diminished when Step 3 is included in the pipeline. Overall, Table 3 shows that the inclusion of phonetic information is consistently beneficial for segmentation in various linguistic scenarios.

We also test our methods on data from another dataset (Dai et al., 2015), in which singers perform three tunes using the syllable /Ta/. In Table 4, we show results for 12 recordings (singers 1,2,4,7). Step 1 performs relatively well because in this case each musical note comprises a voiced segment preceded by a voiceless consonant, so the voicing-based segmentation reflects the note boundaries. In this context Steps 2 and 3 cause split errors and reduce performance.

By analysing specific examples, we identified two sources of errors made by our system. Firstly, there are unrecognized phonemes that lead to merged errors, which could potentially be due to the constraints exerted by the pronunciation and language models of the lyrics transcriber. Secondly, the input pitch track is inactive during voiceless consonants, while the ground truth annotations of the dataset usually include the voiceless consonant at the end of a syllable, resulting in a longer duration note. This disagreement causes a number of offset errors. Despite these errors, we obtained state-of-the-art results on a public dataset for the task of sung note segmentation.

## References


E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. 2013. Automatic music transcription: Challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434.

A. de Cheveigné and H. Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930.

J. Dai and S. Dixon. 2019. Intonation trajectories within tones in unaccompanied soprano, alto, tenor, bass quartet singing. *Journal of the Acoustical Society of America*, 146(2):1005–1014.

J. Dai, M. Mauch, and S. Dixon. 2015. Analysis of intonation trajectories in solo singing. In *16th International Society for Music Information Retrieval Conference*, pages 420–426.

Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2020. Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In *International Joint Conference on Neural Networks*. ArXiv: 2007.06486.

Zih-Sing Fu and Li Su. 2019. Hierarchical classification networks for singing voice segmentation and transcription. In *International Society for Music Information Retrieval Conference*.

Emilia Gómez and Jordi Bonada. 2013. Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to A Cappella singing. *Computer Music Journal*, 37(2):73–90.

Hoon Heo and Kyogu Lee. 2017. Robust singing transcription system using local homogeneity in the harmonic structure. *IEICE Transactions on Information and Systems*, E100.D(5):1114–1123.

Hsuan-Huei Shih, S.S. Narayanan, and C.-C.J. Kuo. 2002. An HMM-based approach to humming transcription. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 337–340. IEEE.

Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. 2015. Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency. In *First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, pages 23–30.

Matthias Mauch and Simon Dixon. 2014. PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663. IEEE.

Emilio Molina, Ana M. Barbancho, Lorenzo J. Tardón, and Isabel Barbancho. 2014. Evaluation framework for automatic singing transcription. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 567–572.

Emilio Molina, Lorenzo J. Tardón, Ana M. Barbancho, and Isabel Barbancho. 2015. SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):252–263.

Matti P. Ryynänen and Anssi P. Klapuri. 2004. Modelling of note events for singing transcription. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*.

Craig Stuart Sapp. 2006. Mazurka Project plugins for Sonic Visualiser. *Poslední aktualizace*, 6(5).

Luwei Yang, Akira Maezawa, Jordan B. L. Smith, and Elaine Chew. 2017. Probabilistic transcription of sung melody using a pitch dynamic model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305. IEEE.