ANIMATED SINGING VOICE

Frank Zalkow[1] **Emir Demirel**[3] **Sebastian Rosenzweig**[1]
Tejaswinee Kelkar[4] **Alejandro Delgado**[2] **Vinod Subramanian** [3]
[1] International Audio Laboratories Erlangen, Germany
[2] ROLI Ltd., London, UK
[3] Queen Mary University of London, London, UK
[4] University of Oslo, Norway

ABSTRACT

In this project, our main aim was to visualize non-score based features of a monophonic singing voice. In particular, we focused on capturing and visualizing note approach style clustering, vibrato, and phonetic residuals (future work). Some of these features, once retrieved, would constitute (or derive in) cues that both composer and singer could benefit when it comes to performance.

1. SIGNIFICANCE STATEMENT

Imagine you are a vocalist trying to understand how exactly what your favorite singer does that isn't well captured in the score, or in the descriptions, here is a tool! (In the making)

Outputs from this project can be used to visualize more-than-score features in the singing voice. For now, this enables you to see where singers have used vibrato, what is the style of approach for each note, and to hear clearly what kinds of phonetic and breath shapes are used for

2. INTRODUCTION

To collect data for this, we started with a jazz standard 'Aint Misbehavin'. We recorded live singing to the score of this song using both a dynamic microphone and a throat microphone, to a reference MIDI track. Using these two types of recordings makes it easier to capture pitch-based vs phonetic features as we demonstrate later. Having recorded singing to the score, we also then obtain annotated events from the midi files, due to which

With the recording we have a midi and a score They are synced so that we have 'ground truth' for what we are starting with

Thereafter, we have extracted pitch Pitch from the score based on the note events

Trying to visualize frame by frame absolute difference between the audio signal and the midi file

3. METHODS

Initially, we have recorded a singing voice performance where the vocal has sung the score of the famous jazz standard 'Ain' Misbehavin' in 100 bpm. We have recorded the singing voice with a throat microphone [5], which is presented by Sebastian Rosenzweig during the first introductory session of the HAMR2018. The reason for using the throat microphone is that its capability of recording the si nging voice according to the throat movements of the singer, minimizing the influence of the noise and artifacts from the outer environment.

The first step in processing the audio is to obtain the pitch track of the singing voice signal. In our approach, we have used an approach [2] similar to the Melodia algorithm [4].

The first visualization we provide is the frame-by-frame pitch deviation of the pitch track from the MIDI notes as seen in Figure 1. In the figure, we have plotted only the regions where there are non-silent MIDI notes are present. It is seen that the vocal performance is closely related to the MIDI notes obtained from the score, with frequent intonation differences. Our goal is to be able to represent these differences and nuances using various non-score type visualizations.

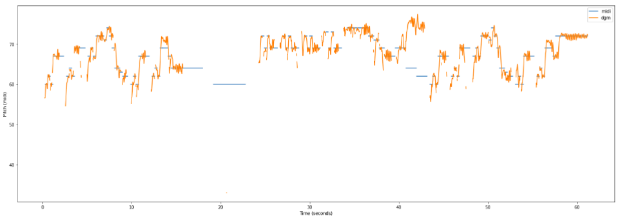

Figure 1. Deviations of the singing voice (vocal) performance from the MIDI notes

In a next step, we segmented the pitch track by the note on- and offset as given by the MIDI file. The shape of a segmented pitch track is then characterized the coefficients of a fifth order polynomial, fitted to that curve. This results in a six-dimensional feature vector for each note.

Attribution: Frank Zalkow, Emir Demirel, Sebastian Rosenzweig, Tejaswinee, Alejandro Delgado, Vinod Subramanian. "Animated Singing Voice", 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

Furthermore the average absolute difference between the segmented pitch track and its corresponding polynomial is taken as a seventh dimension of the feature vector. PCA [1] is then used to project the resulting features vectors into two dimensions, which can be visualized in a scatter plot. When synchronously playing the audio and highlighting the corresponding note in the scatter plot, we that notes with similar characteristics (e.g. vibrato), are located in nearby regions of the plot.

Another aspect of singing voice signals we would like to visualize is the vibrato. In our approach of vibrato detection, we have exploited a method that uses 'Zero Crossing Rate' (as described in [3]). First, we analyze the audio segments of each stable note event separately. For each isolated audio segment, we obtain the median of the frequencies within the segment. Then, the median frequency is set as the threshold for the zero crossing rate algorithm. Then we have set an experimental value of 70 for the threshold to decide if the value of the resulting zero crossing rate indicates a vibrato. The resulting vibrato instances are shown in Figure 2. Even though our system shows sensible results, a further post-processing on the vocal signal needs to be considered to achieve better musicologically meaningful vibrato instances, especially where there are distinct frequency (or pitch) jumps. According to our main goal for visualization, we would like to apply visual clues or color indicators where the system detects there is vibrato in the audio signal.

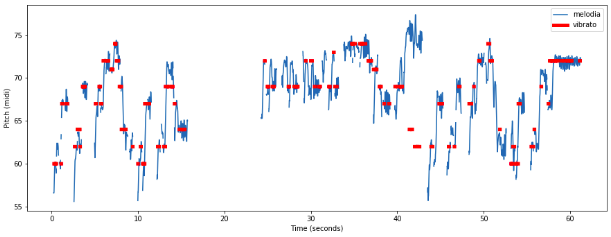

Figure 2. Vibrato instances from the singing voice performance

4. RESULTS AND CONCLUSION

We have developed two systems: the first system uses the features from the deviation of the pitch from the MIDI value to create a visualization and the second is able to detect vibrato in a vocal performance using zero crossing rates which indicates the start and end point of vibratos.

For the first system we experimented with a few different types of features and dimensionality reduction algorithms and determined that using PCA on the polynomial coefficients that approximate the pitch deviations seems to be the most intuitive. Since the performance of this tool depends on a perceptual appreciation of the visualization future work will involve a perceptual evaluation of the different combinations of features and dimensionality reduction algorithms.

For the second algorithm we have verified manually that the algorithm works for the example recording we are using in this project. The algorithm would need to be refined first to make it more robust and then evaluated on a larger dataset to verify its performance. In addition, we need to figure out a more intuitive method to visualize the vibrato.

The ultimate goal of this project is to visualize non-score features with the aim of adding more information to the score. If this technique is successful it would be much easier for performers to capture the original intent of the composer. The best use of this work would be a real time application that follows your voice as you perform and gives you performance notes by way of visualizations.

5. REFERENCES

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Jonathan Driedger and Meinard Müller. Verfahren zur Schätzung der Grundfrequenzverläufe von Melodiestimmen in mehrstimmigen Musikaufnahmen. In Wolfgang Auhagen, Claudia Bullerjahn, and Richard von Georgi, editors, *Musikpsychologie – Anwendungsorientierte Forschung*, volume 25 of *Jahrbuch Musikpsychologie*, pages 55–71. Hogrefe-Verlag, 2015.

[3] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 2004.

[4] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[5] Frank Scherbaum. On the benefit of larynx-microphone field recordings for the documentation and analysis of polyphonic vocal music. In *Proceedings of the International Workshop on Folk Music Analysis*, pages 80–87.