DATA-DRIVEN SOLO VOICE ENHANCEMENT FOR JAZZ MUSIC RETRIEVAL

Stefan Balke[1], *Christian Dittmar*[1], *Jakob Abeßer*[2], *Meinard Müller*[1]

[1]International Audio Laboratories Erlangen, Friedrich-Alexander-Universität (FAU), Germany
[2]Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany

stefan.balke@audiolabs-erlangen.de

ABSTRACT

Retrieving short monophonic queries in music recordings is a challenging research problem in Music Information Retrieval (MIR). In jazz music, given a solo transcription, one retrieval task is to find the corresponding (potentially polyphonic) recording in a music collection. Many conventional systems approach such retrieval tasks by first extracting the predominant F0-trajectory from the recording, then quantizing the extracted trajectory to musical pitches and finally comparing the resulting pitch sequence to the monophonic query. In this paper, we introduce a data-driven approach that avoids the hard decisions involved in conventional approaches: Given pairs of time-frequency (TF) representations of full music recordings and TF representations of solo transcriptions, we use a DNN-based approach to learn a mapping for transforming a "polyphonic" TF representation into a "monophonic" TF representation. This transform can be considered as a kind of solo voice enhancement. We evaluate our approach within a jazz solo retrieval scenario and compare it to a state-of-the-art method for predominant melody extraction.

Index Terms— Music Information Retrieval, Neural Networks, Query-by-Example.

1. INTRODUCTION

The internet offers a large amount of digital multimedia content—including audio recordings, digitized images of scanned sheet music, album covers, and an increasing number of video clips. The huge amount of readily available music requires retrieval strategies that allow users to explore large music collections in a convenient and enjoyable way [1]. In this paper, we consider the retrieval scenario of identifying jazz solo transcriptions in a collection of music recordings, see Figure 1. When presented in a musical theme retrieval scenario for classical music [2], this task offers various challenges, e. g., local and global tempo changes, tuning deviations, or key transpositions. Jazz solos usually consist of a predominant solo instrument (e. g., trumpet, saxophone, clarinet, trombone) playing simultaneously with the accompaniment of the rhythm group (e. g., piano, bass, drums). This typical interaction between the musicians leads to a complex mixture of melodic and percussive sources in the music recording. Consequently, retrieving monophonic pitch sequences of a transcribed solo can be very difficult due to the influence of the additional instruments in the accompaniment.

In this paper, we propose a data-driven approach for enhancing the solo voice in jazz recordings with the goal to improve the retrieval results. As our main technical contribution, we adapt a DNN architecture originally intended for music source separation [3] to train a model for enhancing the solo voice in jazz music recordings. Given the time-frequency (TF) representation of an audio recording as input for the DNN and a jazz solo transcription similar to a piano roll as the target TF representation, the training goal is to learn a mapping between both representations which enhances the solo voice and attenuates the accompaniment.

Throughout this work, we use the jazz solo transcriptions and music recordings provided by the Weimar Jazz Database (WJD). The WJD consists of 299 (as of August 2016) transcriptions of instrumental solos in jazz recordings performed by a wide range of renowned jazz musicians. The solos have been manually annotated and verified by musicology and jazz students at the Liszt School of Music Weimar as part of the Jazzomat Research Project [4]. Furthermore, the database contains more musical annotations (e. g., beats, boundaries, etc.) including basic meta-data of the jazz recording itself (i. e., artist, record name, etc.). A motivation for improving the considered retrieval scenario is to connect the WJD with other resources available online, e. g., YouTube. This way, the user could benefit from the additional annotations provided by the WJD while exploring jazz music.

The remainder of this paper is structured as follows. In Section 2, we discuss related works for cross-modal retrieval and solo voice enhancement approaches. In Section 3, we introduce our DNN-based approach for solo voice enhancement. In particular, we explain the chosen DNN architecture, specify our training strategy, and report on the DNN's performance using the WJD. Finally in Section 4, we evaluate our approach within the aforementioned retrieval scenario and compare it against a baseline and a conventional state-of-the-art system. In our experiments, we show that our DNN-based approach improves the retrieval quality over the baseline and performs comparably to the state-of-the-art approach.

2. RELATED WORK

Many systems for content-based audio retrieval that follow the query-by-example paradigm have been suggested [5–10]. One such retrieval scenario is known as *query-by-humming* [11, 12], where the user specifies a query by singing or humming a part of a melody. Similarly, the user may specify a query by playing a musical phrase of a piece of music on an instrument [13, 14]. In a related retrieval scenario, the task is to identify a short symbolic query (e. g., taken from a musical score) in a music recording [2, 5–7, 15]. Conventional retrieval systems approach this task by first extracting the F0-trajectory from the recording, quantizing the extracted trajectory to musical pitches and finally mapping it to a TF representation to perform the matching (see [12]).

Many works in the MIR literature are concerned with extracting

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen IIS. This work has been supported by the German Research Foundation (DFG MU 2686/6-1).

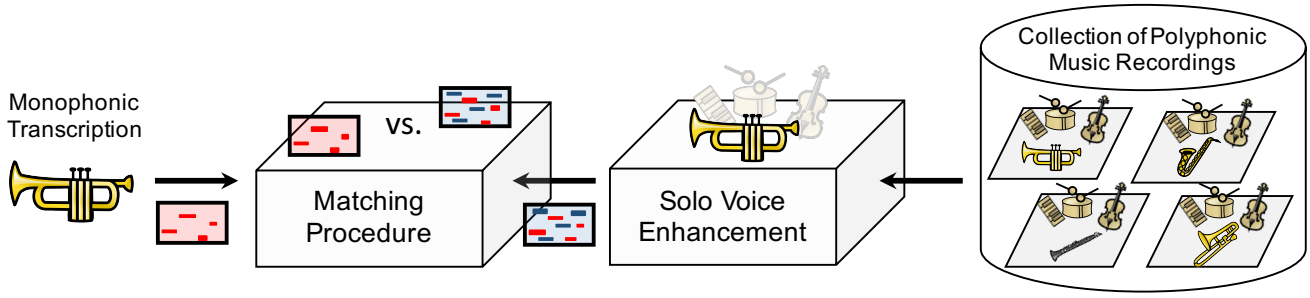

Fig. 1. Illustration of the retrieval scenario. Given a jazz solo transcription used as a query, the task is to identify the music recording containing the solo. By enhancing the solo voice, we reduce the influence of the accompaniment in order to increase the retrieval results.

the predominant melody in polyphonic music recordings—a widely used example is Melodia [16]. More recent studies adapted techniques to work better with different musical styles, e. g., in [17], a combination of estimation methods is used to improve the performance on symphonic music. In [18], the authors use a source-filter model to better incorporate timbral information from the predominant melody source. A data-driven approach is described in [19], where a trained classifier is used to select the output for the predominant melody instead of using heuristics.

3. DNN-BASED SOLO VOICE ENHANCEMENT

Our data-driven solo voice enhancement approach is inspired by the procedure proposed in [3], where the authors use a DNN for source separation. We will now explain how we adapt this DNN architecture to our jazz music scenario.

3.1. Deep Neural Network

Our DNN architecture closely follows [3], where the authors describe a DNN architecture and training protocol for source separation of monophonic instrument melodies from polyphonic mixtures. In principle, the network is similar to Stacked Denoising Autoencoders (SDA) [20], i. e., it consists of a sequence of conventional neural network layers that map input vectors to target output vectors by multiplying with a weight matrix, adding a bias term and applying a non-linearity (rectified linear units). In the setting described by the authors of the original work, the initial DNN consists of 3591 input units, a hidden layer, and 513 output units. The input vectors stem from a concatenation of 7 neighboring frames (513 dimensions each) obtained from a Short Time Fourier Transform (STFT) [21]. The target output vector is a magnitude spectrogram frame (513 dimensions) of the desired ground-truth. The training procedure uses the mean squared error between input and output to adjust the internal weights and biases via Stochastic Gradient Descent (SGD) until 600 epochs of training are reached. Afterwards, the next layer is stacked onto the first one and the output of the first is interpreted as an input vector. This way, the network is gradually built up and trained to a depth of five hidden layers. The originality of the approach in [3] lies in the least-squares initialization of the weights and biases of each layer prior to the SGD training.

In our approach, we do not try to map mixture spectra to solo instrument spectra, but rather to activation vectors for musical pitches. Our input vectors stem from an STFT (frame size $= 4096$ samples, hop size $= 2048$ samples) provided by the `librosa` Python package [22]. We then map the spectral coefficients to a logarithmically spaced frequency axis with 12 semitones per octave and 10 octaves in total which forms the TF representation for the music recordings [21]. The TF representations for the solo transcriptions are directly obtained from the WJD. In these first experiments, we want a simple DNN architecture and do not consider temporal context to keep the number of DNN parameters low. Therefore, our initial DNN consists of 120 input units, one hidden layer with 120 units, and 120 output units. Figure 2 shows the input TF representation of the music recording and the corresponding target output TF representation from the WJD's solo transcription.

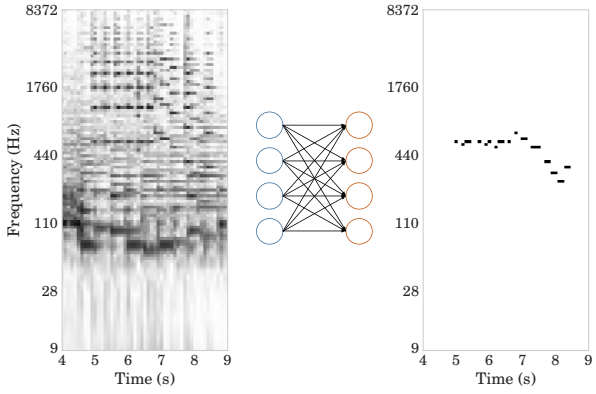

Fig. 2. Input TF representation obtained from a music recording (left) and target TF obtained from the related WJD's solo transcription (right).

3.2. Training

To train our DNNs, we consider the solo sections of the tracks provided by the WJD, i. e., where a solo transcription in a representation similar to a piano-roll is available. This selection leads to a corpus of around 9.5 hours of annotated music recordings. To perform our experiments, we sample 10 folds from these music recordings for training and testing using `scikit-learn` [23]. By using the record identifier provided by the WJD, we avoid using solos from the same record simultaneously in the training and test sets. Furthermore, we randomly split 30 % of the training set to be used as validation data during the training epochs. Table 1 lists the mean durations and standard deviations for the different folds and the portion of the recordings that consists of an actively playing soloist. The low standard deviations in the duration, as well as in the portion of active frames indicate that we created comparable folds. Note that

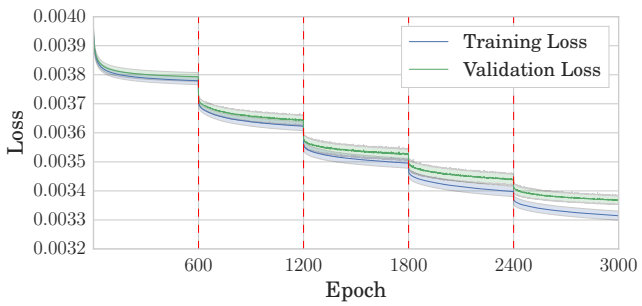

Fig. 3. Training and validation loss during training epochs. For both losses, we show the mean values and the 95 % confidence intervals. The red lines indicate when the next layer is added to the DNN.

	Training Set	Validation Set	Test Set
Duration (h)	5.575 (0.003)	2.389 (0.001)	0.885 (0.004)
Active Frames (%)	61.9 (0.2)	62.0 (0.3)	61.9 (1.8)
No. of Solos	269.1 (5.2)	—	29.9 (5.2)
No. of Full Rec.	204.3 (3.8)	—	22.7 (3.8)

Table 1. Mean duration and mean ratio of active frames aggregated over all folds (standard deviation is enclosed by brackets).

a full recording can contain more than one solo transcription which explains the higher number of solo transcriptions compared to the number of full recordings. In order to reproduce the experiments, we offer the calculated features for all folds, as well as the exact details of the network architecture, on our accompanying website [24].

We start the training with our initial DNN with one hidden layer. We use SGD (momentum $=$ 0.9, batch size $=$ 100) with mean squared error as our loss function. After multiples of 600 epochs, we add the next layer with 120 units to the network until a depth of five hidden layers is reached. All the DNNs have been trained using the Python package `keras` [25]. The resulting mean training and mean validation loss considering all 10 folds are shown in Figure 3. After multiples of 600 epochs, we see that the loss improves as we introduce the next hidden layer to the network. With more added layers, we see that the validation loss diverges from the training loss as a sign that we are slowly getting into overfitting and can thus end the training.

3.3. Qualitative Evaluation

To get an intuition about the output results of the network, we process short passages from solo excerpts with the trained DNNs. Figure 4a shows the TF representation of an excerpt from a trumpet solo. Processing this with the DNN yields the output TF representation as shown in Figure 4b. Note that the magnitudes of the TF representations are logarithmically compressed for visualization purposes. In the output, we can notice a clear attenuation of frequencies below 110 Hz and above 1760 Hz. An explanation for this phenomenon is that no pitch activations in those frequency bands are apparent in our training data. Thus, the DNN quickly learns to attenuate these frequency areas since they do not contribute to the target pitch activations at the output. In the region between these two frequencies, a clear enhancement of the solo voice can be seen, together with some additional noise. As seen in the input TF representation, the fundamental frequency (around 500 Hz) contains less energy than the first harmonic (around 1000 Hz), which is typical for the trumpet.

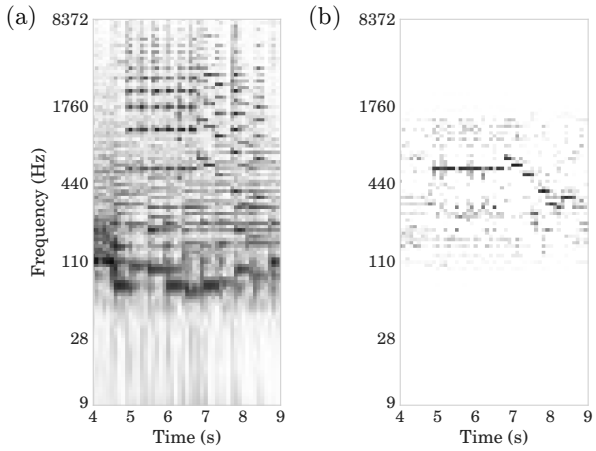

Fig. 4. Typical example for the polyphony reduction using our DNN for an excerpt from Clifford Brown's solo on Jordu. **(a)** Input TF representation. **(b)** Output TF representation after processing with the DNN.

However, the DNN correctly identifies the fundamental frequency. Further examples, as well as sonifications of the DNN's output, can be found at the accompanying website [24].

4. RETRIEVAL APPLICATION

In this section, we first summarize our retrieval procedure and then describe our experiments. We intentionally constrain the retrieval problem to a very controlled scenario where we know that the monophonic queries correspond almost perfectly to the soloist's melody in the recording. We can rely on this assumption, since we use the manual transcriptions of the soloist as provided in the WJD.

4.1. Retrieval Task and Evaluation Measure

In the this section, we formalize our retrieval task following [21]. Let $\mathcal{Q}$ be a collection of jazz solo transcriptions, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be a set of music recordings, which we regard as a database collection consisting of *documents* $D \in \mathcal{D}$. Given a query $Q \in \mathcal{Q}$, the retrieval task is to identify the semantically corresponding documents $D \in \mathcal{D}$. In our experiments, we use a standard matching approach which is based on chroma features and a variant of Subsequence Dynamic Time Warping (SDTW). In particular, we use a chroma variant called CENS features with a smoothing of 9 time frames and a downsampling factor of 2 [26]. Comparing a query $Q \in \mathcal{Q}$ with each of the documents $D \in \mathcal{D}$ using SDTW yields a distance value for each pair (Q, D). We then rank the documents according to the these distance values of the documents $D \in \mathcal{D}$, where (due to the design of our datasets) one of these documents is considered relevant. In the following, we use the mean reciprocal rank (MRR) of the relevant document $D \in \mathcal{D}$ as our main evaluation measure. For the details of this procedure, we refer to the literature, e. g., [21, Chapter 7.2.2].

4.2. Experiments

We now report our retrieval experiments which follow the retrieval pipeline illustrated in Figure 1. In general, for our retrieval experiments, the queries are TF representations of the solo transcriptions

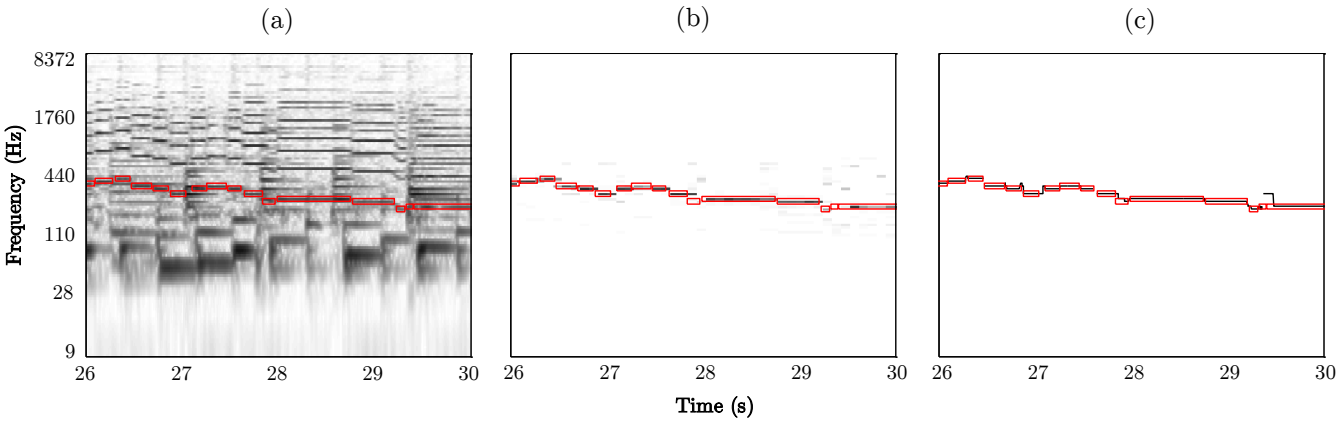

Fig. 5. Typical example for the effect of both solo voice enhancement techniques. **(a)** Log-frequency magnitude spectrogram of a short jazz excerpt from our data. There is a clearly predominant solo melody, but also strong components from the accompaniment, such as bass and drums. **(b)** The same excerpt after running through a trained DNN as described in Section 3. We can see strongly attenuated influence of the accompaniment. **(c)** The same excerpt after extracting the predominant melody using the salience-based approach [16]. We can see that the trajectory of the solo melody has been tracked with only very few spurious frequencies.

from the WJD and the database elements are the TF representations of the corresponding full recordings containing the solos. We perform the retrieval for all 10 training folds separately. As listed in Table 1, the retrieval task consists in average for each fold of 30 solo transcriptions as queries to 23 music recordings in the database. Assuming we have a system that retrieves the relevant document randomly following a uniform distribution, for 30 queries and 23 database elements this would lead to a mean reciprocal rank of 0.13. This value serves as a lower bound of the expected performance of more intelligent retrieval systems. To further study the retrieval robustness, we consider query lengths starting from using the first 25 s of the solo transcription and then successively going down to 3 s.

In our baseline approach, we reduce the TF representations of the query and database documents (without using the DNN) to chroma sequences and apply the retrieval technique introduced earlier. The results of the baseline approach in terms of MRR for different query lengths are shown in Figure 6, indicated by the blue line. For a query length of 25 s, the baseline approach yields an MRR of 0.94. Reducing the query length to 5 s leads to a significant drop of the MRR down to 0.63. Now we consider our proposed DNN-based solo voice enhancement approach. The queries stay the same as in the baseline approach, but the TF representations of the database recordings are processed with our DNN before we reduce them to chroma sequences. For a query length of 25 s, this yields an MRR of 0.98; for a query length of 5 s, the MRR only slightly decreases to 0.86 which is much less than in the baseline approach. A reason for this is that the queries lose their specificity the shorter they become. This leads to wrong retrieval results especially when using the unprocessed recordings as in the baseline approach. The DNN-based approach compensates this by enhancing the solo voice and therefore makes it easier for the retrieval technique to identify the relevant recording.

Lastly, we consider a salience-based approach described in [16] for processing the music recording's TF representation. In short, this method extracts the predominant melody's F0-trajectory from the full recording, which is then quantized and mapped to a TF representation. The conceptional difference to our DNN-based approach is illustrated in Figure 5. For a query length of 25 s, this method yields a slightly lower MRR than the DNN-based approach of 0.96. Reducing the query to a length of 5 s, we achieve an MRR of 0.84. All three methods perform well when considering query lengths of more than 20 s. When the query length is shortened, all methods show a decrease in performance, whereas the DNN-based and salience-based methods significantly outperform the baseline approach.

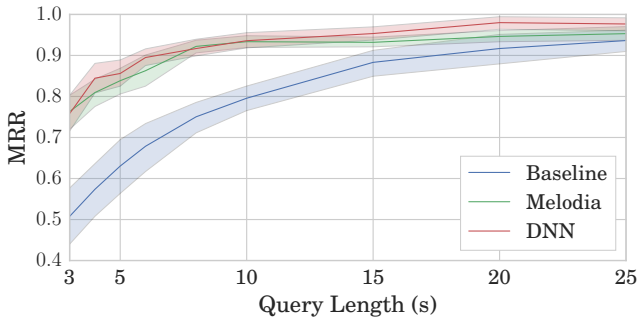

Fig. 6. Mean reciprocal rank (MRR) for all three methods performed on all folds and with varying the query length. For all methods, we show the 95 % confidence intervals.

5. CONCLUSION

In this paper, we described a data-driven approach for solo voice enhancement by adapting a DNN-based method originally used for source separation. As a case study, we used this enhancement strategy to improve the performance of a cross-modal retrieval scenario and compared it to a baseline and a conventional method for predominant melody estimation. From the experiments we conclude that in the case of jazz recordings, solo voice enhancement improves the retrieval results. Furthermore, the DNN-based and salience-based approaches perform on par in this scenario of jazz music and can be seen as two alternative approaches. In future work, we would like to investigate if we can further improve the results by enhancing the current data-driven approach, e. g., by incorporating temporal context frames or testing different network architectures.

6. REFERENCES

[1] Cynthia C. S. Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *Proc. of the Int. ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM)*, 2011, pp. 1–6.

[2] Stefan Balke, Vlora Arifi-Müller, Lukas Lamprecht, and Meinard Müller, "Retrieving audio recordings using musical themes," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 281–285.

[3] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 2135–2139.

[4] The Jazzomat Research Project, "Database download, last accessed: 2016/02/17," `http://jazzomat.hfm-weimar.de`.

[5] Christian Fremerey, Michael Clausen, Sebastian Ewert, and Meinard Müller, "Sheet music-audio identification," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009, pp. 645–650.

[6] Jeremy Pickens, Juan Pablo Bello, Giuliano Monti, Tim Crawford, Matthew Dovey, Mark Sandler, and Don Byrd, "Polyphonic score retrieval using polyphonic audio," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, France, 2002.

[7] Iman S.H. Suyoto, Alexandra L. Uitdenbogerd, and Falk Scholer, "Searching musical audio using symbolic queries," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 372–381, 2008.

[8] Michael A. Casey, Remco Veltkap, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[9] Peter Grosche, Meinard Müller, and Joan Serrà, "Audio content-based music retrieval," in *Multimodal Music Processing*, Meinard Müller, Masataka Goto, and Markus Schedl, Eds., vol. 3 of *Dagstuhl Follow-Ups*, pp. 157–174. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

[10] Rainer Typke, Frans Wiering, and Remco C. Veltkamp, "A survey of music information retrieval systems," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, London, UK, 2005, pp. 153–160.

[11] Matti Ryynänen and Anssi Klapuri, "Query by humming of MIDI and audio using locality sensitive hashing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 2249–2252.

[12] Justin Salamon, Joan Serrà, and Emilia Gómez, "Tonal representations for music retrieval: from version identification to query-by-humming," *Int. Journal of Multimedia Information Retrieval*, vol. 2, no. 1, pp. 45–58, 2013.

[13] Andreas Arzt, Sebastian Böck, and Gerhard Widmer, "Fast identification of piece and score position via symbolic fingerprinting," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Porto, Portugal, 2012, pp. 433–438.

[14] Taro Masuda, Kazuyoshi Yoshii, Masataka Goto, and Shigeo Morishima, "Spotting a query phrase from polyphonic music audio signals based on semi-supervised nonnegative matrix factorization," in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2014, pp. 227–232.

[15] Colin Raffel and Daniel P. W. Ellis, "Large-scale content-based matching of MIDI and audio files," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Málaga, Spain, 2015, pp. 234–240.

[16] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[17] Juan J. Bosch, Ricard Marxer, and Emilia Gómez, "Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music," *Journal of New Music Research*, vol. 45, no. 2, pp. 101–117, 2016.

[18] Juan J. Bosch, Rachel M. Bittner, Justin Salamon, and Emilia Gómez, "A comparison of melody extraction methods based on source-filter modelling," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, New York City, USA, 2016, pp. 571–577.

[19] Rachel M. Bittner, Justin Salamon, Slim Essid, and Juan Pablo Bello, "Melody extraction by contour classification," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Málaga, Spain, 2015, pp. 500–506.

[20] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, Helsinki, Finland, June 2008, pp. 1096–1103.

[21] Meinard Müller, *Fundamentals of Music Processing*, Springer Verlag, 2015.

[22] Brian McFee, Matt McVicar, Colin Raffel, Dawen Liang, Oriol Nieto, Eric Battenberg, Josh Moore, Dan Ellis, Ryuichi Yamamoto, Rachel Bittner, Douglas Repetto, Petr Viktorin, João Felipe Santos, and Adrian Holovaty, "librosa: 0.4.1," 2015.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] Stefan Balke, Christian Dittmar, and Meinard Müller, "Accompanying website: Data-driven solo voice enhancement for jazz music retrieval," `http://www.audiolabs-erlangen.de/resources/MIR/2017-ICASSP-SoloVoiceEnhancement/`.

[25] François Chollet, "Keras," `https://github.com/fchollet/keras`, 2015.

[26] Meinard Müller, Frank Kurth, and Michael Clausen, "Chroma-based statistical audio features for audio matching," in *Proc. of the Workshop on Applications of Signal Processing (WASPAA)*, New Paltz, New York, USA, 2005, pp. 275–278.