FILOSAX: A DATASET OF ANNOTATED JAZZ SAXOPHONE RECORDINGS

Dave Foster
Queen Mary University of London
d.foster@qmul.ac.uk

Simon Dixon
Queen Mary University of London
Centre For Digital Music

ABSTRACT

The Filosax dataset is a large collection of specially commissioned recordings of jazz saxophonists playing with commercially available backing tracks. Five participants each recorded themselves playing the melody, interpreting a transcribed solo and improvising on 48 tracks, giving a total of around 24 hours of audio data. The solos are annotated both as individual note events with physical timing, and as sheet music with a metrical interpretation of the timing. In this paper, we outline the criteria used for choosing and sourcing the repertoire, the recording process and the semi-automatic transcription pipeline. We demonstrate the use of the dataset to analyse musical phenomena such as swing timing and dynamics of typical musical figures, as well as for training a source activity detection system and predicting expressive characteristics. Other potential applications include the modelling of jazz improvisation, performer identification, automatic music transcription, source separation and music generation.

1. INTRODUCTION

The study of jazz improvisation has often focused on modelling *what* to play, most recently via deep learning techniques such as transformers [1], language models [2] and GANs [3]. Other recent work [4] has suggested that *how* to play is of equal importance when generating convincing synthesised performances. To properly examine the minutiae of how a performer plays, one requires a wealth of clean, isolated and consistent recordings, which are hard to come by when looking specifically at jazz music.

Another issue when researching the expressive nature of jazz performances is the difficulty of making pair-wise comparisons between performers. This is because there is very little overlap between the recorded corpora of any two musicians, especially when hoping for consistency of both key and tempo. Whereas classical music researchers can draw upon multiple recordings of the same pieces [5, 6], jazz music researchers have only sparse instances of duplicated "head" statements and common "licks" upon which to make their comparisons.

These were the motivations behind the commissioning and curation of the Filosax dataset, which was designed to provide both the isolated recordings and homogeneity of stimuli to allow for the analytical fidelity and inter-participant consistency that are required. The period 2020-2021 inadvertently proved a good period for enrolling willing participants, as COVID-related lockdowns meant that there was a dearth of live performing opportunities, and hence expert performers and improvisers had more time and inclination to take part than they might otherwise have done. The downside of the lockdown backdrop was that we were unable to record the musicians in exactly the same environment. We attempted to mitigate for this by collating an environment-agnostic recording kit, which was sent between the participants and meant that the recordings are as consistent as possible.

A novel aspect of the dataset is the dual note-level annotations that accompany the audio data. The first is a segmentation of the soloist audio into discrete note-events with pitch tags and precise timings, which allows the user to determine exactly *how* a note was played. The precise perturbations of pitch, amplitude and timbre can be measured and quantified, unencumbered by the requirement to filter out (or interpolate from) the accompaniment. The second is a sheet music representation (at a level of detail akin to the "Omnibook" series [1]), where each note is assigned a place in the metrical grid by an expert human jazz transcriber. The mapping between the two annotation layers provides, perhaps, the most useful insight: given a sequence of notes, how does the performer play each one, given its position in the sequence?

We recognise that a set of annotated, source separated recordings will also be of use to researchers in related fields, some of which are discussed in section 7. Due to licensing issues, the full Filosax dataset cannot be made publicly available, but suitable researchers are welcome to apply to the authors to receive copies of the non-copyright material and annotations, along with instructions on how to purchase the copyrighted material and automatically reconstruct the full dataset.

2. RELATED WORK

The Weimar Jazz Database (WJD) [7] was a landmark in annotated jazz data, with 456 manually transcribed solos by 78 performers, and featuring music from a broad range

[1] https://www.halleonard.com/series/OMNIBK?subsiteid=65&&dt=item#products

of eras (1925-2009). The inclusion of additional information about idiom, style, and form allow it to be used for many MIR purposes as well as ethnomusicological research. For our purposes, it is hampered by the medium (stereo mixes) and original choice of repertoire (chosen by the performers for artistic rather than scientific reasons) from which the authors could only take samples. There are only 3 pieces played by multiple tenor saxophonists ("Body and Soul" by 12 musicians, "Night and Day" and "U.M.M.G." both by 2 musicians), and there are no readable (human corrected) score representations.

The "Dig That Lick" project developed pattern mining within the WJD [8] as well as the collation of the DTL1000 dataset [9], a set of automated transcriptions of 1750 solos from 1060 tracks. The algorithm used for transcribing the melody/lead line achieves a mean F1 score of 0.85, which the authors suggest is adequate for large-scale pattern mining, but implies that a significant amount of manual correction would needed for the data to be of sufficient accuracy for note-level analysis.

The MedleyDB database [10] contains almost 200 tracks with full mixes and separate instrumental/vocal stems in a variety of styles and with pitch annotations for the lead line. Less than 10, however, are appropriate jazz recordings, representing approximately 15 minutes of data.

The best practices for the definition and presentation of an MIR dataset [11, 12] were useful in guiding the design and presentation of our dataset. The processes of compiling and annotating the DALI dataset [13] (of synchronised note/lyric annotations) were similar to the approaches that we have used, as is the means of addressing the issues around the distribution of copyrighted material.

3. DATASET CURATION

The diversity within the corpus of recorded jazz music is so broad that an attempt to capture the full variety of it within a database of this size would be futile. We chose to set a goal of a focusing on depth over breadth and, to this end, decided that we would record a single instrument (the tenor saxophone, possibly the most ubiquitous melodic instrument in recorded jazz) and a narrow scope of mid-tempo "standards" with a quasi-fixed tempo, in 4/4 time and a "swing" feel. We chose to engage five expert performers and improvisers on the instrument, to capture a variety of expressive approaches and to allow for inter-participant comparisons to be made.

3.1 Repertoire

To collect a sufficiently varied set of notes, whilst capturing sufficient repeated elements, we decided to base the dataset on a representative repertoire of pieces. Each participant would record themselves playing these pieces with the same accompaniment, and on each piece would play the melody (the "head"), interpret a transcribed solo, and improvise their own solo.

3.1.1 Accompaniment

A dataset which fully encompasses the jazz improvisation process would capture the interaction between the soloist and rhythm section. The Filosax dataset does not attempt to capture this, as to do so would sacrifice the comparisons which it allows to be made between the various performances under identical conditions. Hence, the interaction process is only in one direction: the soloist can respond to what is heard on the accompaniment, but there is no opportunity for the accompaniment to respond or for a feedback loop to be established.

We chose to use pre-recorded accompaniments from Jamey Aebersold [2], a commercial library of "play-along" tracks recorded between 1967 and the present day. The library consists of over 1000 tracks of jazz standards, recorded with different musicians but with a similar audio presentation of piano+drums in the left channel and bass+drums in the right channel. The use of commercial recordings for this purpose greatly extended the potential repertoire from which we could choose: the alternatives were to use freely available resources (of which there are very few), commission more recordings or to use synthesised accompaniments. Using any of the alternatives would be to the detriment of the range or quality of the data, or would require much more sophisticated recordings. The downside to using the commercial recordings is that we will be unable to distribute them with the dataset.

3.1.2 Solo transcriptions

For the transcribed solos, we sought a group of celebrated jazz artists, whose stylistic output was somewhat similar to each other. We selected 6 such musicians who met the following criteria:

- Made recordings in the 1950's, 1960's and 1970's,
- Made recordings with a discernible and repeated chord sequence,
- The set of their recorded corpus intersects with the set of available backing tracks (in the same key and at a similar tempo).

The musicians chosen (who could broadly be described as performing within the "hard-bop" sub-genre) were: Stan Getz, Dexter Gordon, Tubby Hayes, Joe Henderson, Sonny Rollins and Ben Webster. Each is represented in the dataset by 8 extracts of their recorded solos, from the private collections of the authors, which were transcribed and typeset by the authors. The details of the original recordings were made available to participants, in case they felt it would be useful in developing their own interpretation.

3.1.3 Piece choice

All of the available Jamey Aebersold accompaniment tracks were examined and meta-data extracted (tempo, key, duration, number of choruses, time signature, rhythmic feel). Candidate pieces were found using the following criteria:

[2] `http://jazzbooks.com/jazz/JBIO`

- Entirely in 4/4 time, with a quasi-fixed tempo (allowing for gradual changes),
- With a "swing" feel throughout,
- A repeated chord sequence with sufficient harmonic movement to not be classed as "modal" (which we define as having multiple sections where a single chord is held for longer than 4 bars),
- Tempo in the range 100-240 beats per minute,
- An accompaniment consisting exclusively of piano, bass and drums.

This truncated list was cross-referenced with the discographies of the jazz artists listed in section 3.1.2 where, for a piece to be considered, the recorded version must be: in the same key, within 10% of the backing track tempo, and with at least 1 "chorus" of improvisation by the artist. When a selection (larger than 8) was available, a broad range of tempos, keys and modalities was sought. The full list of pieces is given on the dataset web page [3] .

3.2 Participants

The five musicians recruited to make the recordings were all known to the authors as expert performers and improvisers on the tenor saxophone. They were invited to participate on the understanding that they had access to an appropriate space in which to record themselves, a good quality instrument and a computer capable of making the recordings. On completion of the recordings, they were asked to take part in an informal follow-up session to review the transcriptions and to reflect on their experiences of making the recordings. Each read and agreed to the ethics, information and participation forms associated with the study, and was compensated with an Amazon gift voucher for £100 on completion of their recordings.

3.3 Recording

The recordings were made consecutively by the five participants, in their own homes, having been supplied with a set of materials, recording equipment and instructions.

3.3.1 Materials

For each of the 48 pieces, the participants were supplied with a printed copy of the sheet music and a copy of the corresponding digital audio workstation (DAW) file (segmented into bars and choruses) to record into. Access to the materials was given prior to receiving the recording equipment, in case they wanted to prepare.

3.3.2 Equipment

A flight case of equipment was shipped between the participants, containing an Aston Stealth microphone, a Focusrite Scarlett Solo USB audio interface, a reflection shield, closed-back headphones, microphone stand, XLR cable and USB cable. Directions on how to assemble and set up the equipment were given, where the settings for the microphone and audio interface were prescribed. Suggestions of how to choose and prepare a suitable room were included, as well as the exact positioning of the microphone from the instrument.

3.3.3 Instructions

Participants received a document containing detailed information on the goals of the data collection, and how to approach their interpretation of the material. For the "head" section, they were told to perform this freely, as if in a live performance: adding, removing, changing or moving notes as they please (whilst still ensuring that it is identifiably the melody). For the interpretation of the recorded solo, they were asked to play accurately, to the best of their ability, without intentionally changing notes but with grace notes, articulation, slurs and "scoops" as they felt. For the improvised sections, they were asked to approach this more as a practice session than a concert: that is, to include repetition, longer notes, and to explore the full tessitura and dynamic range of the instrument, more than they might otherwise do in a concert setting.

4. ANNOTATION

The recordings in their entirety were annotated at two levels: firstly, an accurate list of note start times, end times, and homogenised pitch (semitone granularity); secondly, a simplified sheet music representation, where an interpretation of the intended rhythm was notated, using a pre-determined set of granularity assumptions. Figure 1 shows the semi-automatic transcription pipeline that was developed for the expedited and accurate annotation of the recordings. Commercially available software was used to aid with the rough segmentation of the audio into discrete note events, before standard MIR packages were used to obtain the absolute temporal boundaries of each event.

4.1 Initial Annotation

4.1.1 Bar / Beat Annotation

The annotation process began before the recordings were made. The backing track audio files were imported into the DAW Logic [4] , having first been normalised into the range $[-1, 1]$ in Audacity. The "Smart Tempo" function (a proprietary beat mapping algorithm) was used to annotate the file into bars and beats, and corrected by the authors where the downbeat or tempo scale was wrongly inferred. Choruses were duplicated or deleted at this stage, so that the duration of each piece was between 5:50 and 6:20 minutes. Finally, markers were added to correspond with the rehearsal marks on the sheet music, to visually guide the participants whilst they were recording.

4.1.2 Approximate note segmentation

When a participant returned a completed file, the note segmentation transcription could take place. This was initially done with the Logic "Flex Pitch" function (another proprietary algorithm, which performs pitch and onset detection), with human correction both before and after conversion to MIDI. The authors found that, on average, 9%

[3] https://dave-foster.github.io/filosax/

[4] https://www.apple.com/uk/logic-pro/

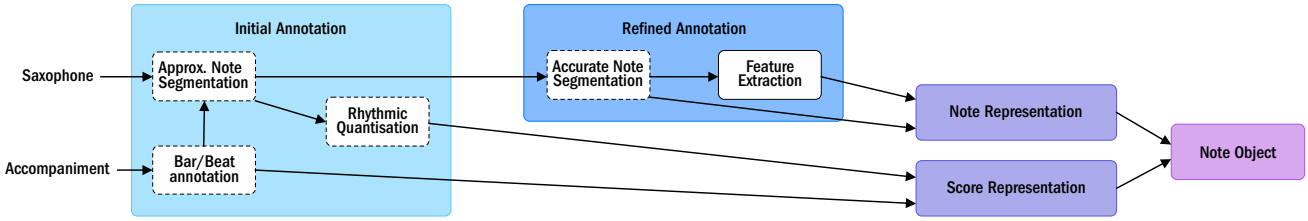

Figure 1: The annotation pipeline. Blocks with dashed lines signify those steps which are manually checked.

of notes in a file needed revising for pitch, timing, separation or concatenation. At this stage, the ground truths regarding the correct number (and order) of note events, their homogenised pitches and approximate timings were established, with the exact values being confirmed at a later stage, described in section 4.2. This workflow was found to maximise the utility of the human intervention (identifying pitches and approximate note boundaries), allowing the exact timings to be sought automatically.

4.1.3 Sheet music representation

The resulting MIDI track from the previous stage was duplicated for the basis of the sheet music representation, which was found to be of sufficient accuracy for this purpose. A granularity of triplet semiquavers was used, but only for notes which were as short as that (for example, in a rapid run of notes). In the main, we determined that a quantisation to a grid of quavers (for both onsets and offsets) would best match the idiomatic style of notation and, by doing so, adhere to the convention of notating off-beat quavers as being "on the grid", regardless of the amount of "swing" that was used to delay them. After this initial quantisation, notes of a shorter duration were human-edited for rhythmic value on a case-to-case basis, where an ethos of readability over absolute temporal accuracy was employed in ambiguous cases.

Grace notes were consistently notated as acciaccatura, placed a triplet semiquaver ahead of the beat, for easier identification later in the pipeline. No slurs, dynamics or articulations are included in the annotations, which are left by the authors for possible future rounds of annotation, where the audio data could be used to generate suggestions in each case. In the sheet music which accompanies the dataset, there is a small exception, in that staccato crochets are used for readability whenever an on-the-beat quaver is followed by a quaver rest (similarly for staccato quavers).

4.2 Refined Annotation

4.2.1 Note boundaries

The three audio files (saxophone and split backing track) were exported from Logic, along with a MIDI file, containing the bar and beat timings and the two sets of note annotations. These were all processed by a Python script (using mido [5] for MIDI functions), for finding accurate note timings and for serialising the data into the final database format. The former was performed by separating the audio into "phrases" (consecutive sequences of notes found in the sheet music representation) by identifying gaps between note events in the score representation. These "phrases" may not correspond with the traditional definition, as they can contain just a single note.

The time values from the approximate note segmentation were used as estimates of the temporal mid-point of each note, so for a note N_k with approximate start and end times $\hat{t}^s_k$ and $\hat{t}^e_k$, the mid-point $\hat{t}^m_k = (\hat{t}^s_k + \hat{t}^e_k)/2$.

We formally describe the process for refining the note boundaries as follows. Each performer's interpretation of a piece is said to consist of a sequence of I phrases $(P_i)_{i=1}^{I}$, where each phrase P_i is a sequence of K_i sounding entities N_k with approximate start, mid-point and end ($\hat{t}^s_k$, $\hat{t}^m_k$ and $\hat{t}^e_k$), start time t^s_k, end time t^e_k and pitch f_k. Hence,

$$P_i = (N_k)_{k=1}^{K_i}, \quad \text{where } N_k = (\hat{t}^s_k, \hat{t}^m_k, \hat{t}^e_k, t^s_k, t^e_k, f_k). \tag{1}$$

The phrases do not overlap, so $P_i(t^e_{K_i}) \leq P_{i+1}(t^s_1)$ for each i. Iterating by phrase, the corresponding audio was analysed using both the Madmom [14] "CNNOnsetProcessor" and Essentia [15] PYIN [16] implementation, giving a sequence of onsets $(O_j)_{j=1}^{J}$, a pitch curve $F(t)$ (a sequence of real or null values for each time frame $t \in [P_{i-1}(t^e_k), P_{i+1}(\hat{t}^s_1))$ and a loudness curve $L(t)$ (a sequence of loudness values for each time frame t).

The start time t^s_k and end time t^e_k (in frames, where $t^s_k < \hat{t}^m_k < t^e_k$) for each N_k is determined by looking backward and then forward from the mid-point, so,

$$t^s_k = \max\{t^e_{k-1}, F^s_{null} + 1, L^s_{quiet} + 1\}, \tag{2}$$

$$t^e_k = \min\{F^e_{null} - 1, O_{first}, L^e_{quiet}\}, \tag{3}$$

where F^s_{null} and F^e_{null} are the time steps of the first null pitch value encountered, counting backwards and forwards from the mid-point $\hat{t}^m_k$ respectively, O_{first} is the first onset encountered after the mid-point, and L^s_{quiet} and L^e_{quiet} are the first times (before and after the mid-point) when the loudness $L < L_{thresh}$, a threshold value. The notes are sequential and monophonic, hence $t^e_k \leq t^s_{k+1}$ for each k.

In the event of a continuous pitch curve and absence of a detected onset on or around the expected position of the boundary between two entities N_k and N_{k+1}, we define:

$$t^e_k = \begin{cases} \hat{t}^e_k, & \text{when} f_k = f_{k+1}, \\ \text{argmin}_t\{F(t) > \dfrac{f_k + f_{k+1}}{2}\}, & \text{when} f_k < f_{k+1}, \\ \text{argmin}_t\{F(t) < \dfrac{f_k + f_{k+1}}{2}\}, & \text{when} f_k > f_{k+1}. \end{cases} \tag{4}$$

Where an onset occurs before the first pitch value (likely due to a breath or key noise) or there is a gap in the pitch

[5] https://github.com/mido

curve between neighbouring notes ($t^e_k < t^s_{k+1}$), an unpitched entity object is recorded in the sequence. If K^P_i is the number of pitched entities and K^U_i the number of unpitched entities in each phrase P_i, then $K_i = K^U_i + K^P_i$ and $K^P_i \geq K^U_i$.

The output of this process is manually validated by the annotator by listening concurrently to the original recording and a synthesised version of the annotations.

4.2.2 Note attributes

With the exact timing of all the notes now known, the attributes of each note can be extracted. Another Python script is employed to automate this process, where it captures both the continuous curves and interpolates landmark values. The pitch curve is extracted (again with the Essentia PYIN function, and constrained to the vicinity of the determined pitch), and the average pitch, time to average pitch, and average vibrato rate and extent (using Essentia functions) are all estimated. Similar curves and features are extracted for amplitude, spectral centroid and spectral flux. It is these attributes that we identified as crucial to the initial goal of studying expressive performance: other use cases for the dataset may require different features, the extraction of which could be automated in a similar fashion. The chord(s) over which the note is played is derived from the published chord sequence, and used to derive the scale degree(s) relative to the chord root.

4.3 Dataset structure

The dataset D is presented as an ordered set of uniquely identifiable sounding entities (pitched and unpitched), which have the following attributes:

- `start_time`,
- `end_time`,
- `musician_number`,
- `piece_number`,
- `bar_number`,
- `bar_type` (head, written solo or improvisation),
- `tempo`.

In addition, pitched entities have the following attributes:

- `MIDI_pitch`,
- `score_start_time`,
- `score_end_time`,
- `score_rhythmic_position`,
- `score_rhythmic_duration`,
- `is_grace_note`,
- `chord_changes` (an array, of length 1 for short notes, and possibly > 1 for longer notes),
- `scale_degrees` (an array).

The entities are collected sequentially in a JSON file (conforming to the JAMS specifications [17]), allowing for easy searching, analysis and n-gram construction. The data is also made available as a set of both MIDI and MusicXML files (the latter just containing the sheet music representation), although the entities themselves contain all the information needed to reconstruct both of these formats. The attributes described in section 4.2.2 are also contained in the JSON file, as are the corresponding full curve values.

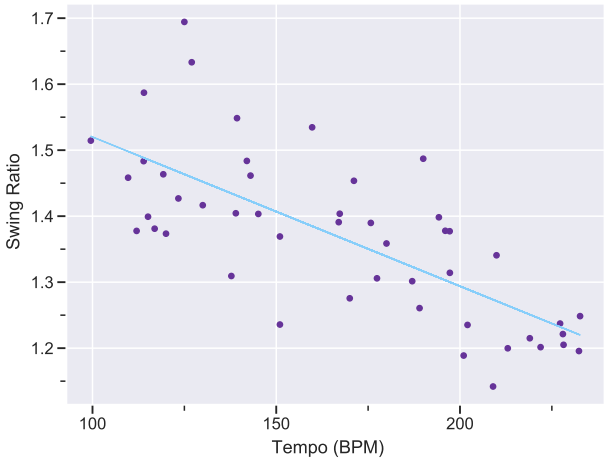

Figure 2: The 'swing ratio' of a single participant as a function of tempo. Each point represents the mean tempo and ratio of quavers played in a single piece.

5. ANALYSIS

We present the results of two analysis studies of the Filosax dataset, making comparisons with previous analysis of similar data, as demonstrations of how the dataset can be used.

5.1 Swing Ratio

The presentation of the pitched entities (described in section 4.3) allows for the "groove" or "swing" of a note to be instantly calculated, by comparing the `score_start_time` attribute (the time at which the note starts in the score representation) to the `start_time` attribute (when the note actually starts).

Figure 2 shows the "swing ratio", the duration of the first of a pair of quavers divided by the duration of the second, as a function of the current speed of the piece. The blue line-of-best-fit shows the same negative correlation found in other analyses of swing rhythm [18–20].

5.2 Enclosed notes

"Enclosing" notes is a device used in the construction of "bebop" phrases, where a chord tone is preceded by both a note above and below (diatonically or chromatically). To show how the dataset can be used for deriving performance parameters, we search the dataset for instances of these 3-grams of consecutive quavers.

Figure 3 shows the range of loudness values where the third note is an off-beat chord tone, and the preceding notes are above and then below that pitch. The graph on the left is derived from instances where one or both of the preceding notes are greater than a tone away, and on the right where both preceding notes are within a tone ("enclosed").

In the first case, the notes are given almost equal emphasis, whereas the second case shows a characteristic emphasis on the first and (to a lesser extent) third notes, by means of "ghosting" (placing less emphasis on) the second. The ranges derived could be used as probability distributions for generating phrase-level performance instructions.

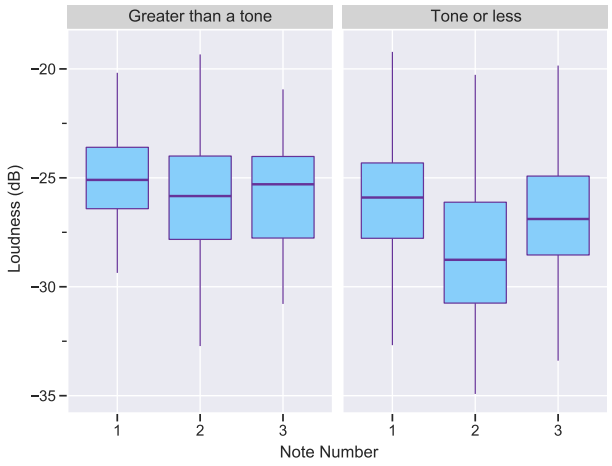

Figure 3: Relative loudness of quavers approaching an off-beat chord tone (3) from above (1) and below (2).

6. EXPERIMENTS

6.1 Activity detection

The Filosax dataset was used to train a system for saxophone activity detection from mixed recordings, that outputs a binary value for each time frame. The solo saxophone recordings were used to generate the ground truth, and mixes of the three stems were used as input. A variation of the U-Net architecture used for the similar task of vocal detection [21] was employed, and various input representations were experimented with.

The best performing combination used a CQT input (achieving an AUROC mean of 0.933), slightly higher than was achieved for vocal activity detection in the original paper (AUROC mean of 0.924). Neither the use of joint training (separation and detection), HCQT input nor HCQT with phase input yielded any improvement in results.

6.2 Expressive Timing

A sequence-to-sequence language model was used to predict performance parameters by framing the problem as a translation task: a "sentence" of "words" (a phrase of notes, using attributes from the dataset) was "translated" into a sequence of expressive instructions (timing, loudness and articulation, all derived from the dataset) via word embeddings and a context vector.

This preliminary system was able to learn the fundamentals of "swing" rhythm (despite it not being explicitly encoded in the input representation), but the rendered output, in its current state, is not of a standard that will impact the bookings of any human jazz musicians. Refinement of this system will form the basis of future research.

7. POTENTIAL APPLICATIONS

We outline several potential applications for the Filosax dataset, outside the field of expressive performance.

7.1 Jazz Improvisation

The dataset contains multiple hours of improvised jazz solos, with corresponding chord and form annotations, which those researching *what* to play could use to train their systems, either as a standalone resource or augmenting another dataset. The range of performances of the "head" of each piece could aid in the study of melody interpretation, and the performances of the transcribed solo could inform research in jazz education.

7.2 Predominant Melody Extraction

Predominant melody extraction in jazz has been restricted to using note annotations with the ensemble recording source [22, 23]. With the Filosax dataset, this type of system could be trained with various mixes of soloist/accompaniment, potentially leading to a system which is more robust to variations in melodic prominence.

7.3 Source Separation

Similar to the previous use case, the 3 distinct audio stems could be leveraged to develop jazz-specific source separation architectures, or existing architectures could be trained with the data. This could lead to improved methods for isolating the solo instrument on jazz recordings which, in turn, could inform more accurate automated data collection from ensemble recordings.

8. DISTRIBUTION

The backing tracks and the melodies are all under copyright, so the Filosax dataset cannot be made public. To ensure reproducibility and to facilitate the adoption of the dataset, we will allow researchers (on application) to access the saxophone recordings, annotations and sample notebooks on the Zenodo repository [6]. We will also provide the list of backing tracks required, and a Python module for checking, normalising and segmenting (see section 4.1.1) the files, in order to accurately reconstruct the data. The module is part of *mirdata* [24], an open-source tool for the distribution of datasets and corresponding annotations, which ensures that the user has the canonical version of all the components.

9. CONCLUSION

No jazz musician ever decided the programme of their concert or album based on what might be useful to future scientists, nor intentionally played an identical chorus to that played by one of their peers because it would provide useful data. We propose that the introduction of the Filosax dataset somewhat tackles these issues, and does so without unduly compromising the stylistic or artistic credibility of the music. The data has already proved to be an invaluable resource for our ongoing research, and we share our rationale, methodology and the data itself in the hope that it may also be for others.

[6] `https://zenodo.org`

10. ACKNOWLEDGMENTS

The first author is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1].

11. REFERENCES

[1] S.-L. Wu and Y.-H. Yang, "The jazz transformer on the front line," in *21st International Society for Music Information Retrieval Conference*. ISMIR, 2020.

[2] S. H. Hakimi, N. Bhonker, and R. El-Yaniv, "Bebopnet: Deep neural models for personalized jazz improvisations," in *21st International Society for Music Information Retrieval Conference*. ISMIR, 2020.

[3] N. Trieu and R. Keller, "JazzGAN: Improvising with generative adversarial networks," in *Proc. Int. Workshop on Musical Metacreation*, 2018.

[4] K. Frieler and W.-G. Zaddach, "Evaluating an analysis-by-synthesis model for jazz improvisation," 2021, under review.

[5] K. Kosta, O. F. Bandtlow, and E. Chew, "Mazurkabl: Score-aligned loudness, beat, expressive markings data for 2000 Chopin Mazurka recordings," in *Proceedings of the Fourth International Conference on Technologies for Music Notation and Representation (TENOR)*, 2018, pp. 85–94.

[6] T. Gadermaier and G. Widmer, "A study of annotation and alignment accuracy for performance comparison in complex orchestral music," in *20th International Society for Music Information Retrieval Conference*. ISMIR, 2019.

[7] M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart, *Inside the Jazzomat: New Perspectives for Jazz Research*. Schott Music GmbH, 2017.

[8] K. Frieler, F. Höger, M. Pfleiderer, and S. Dixon, "Two web applications for exploring melodic patterns in jazz solos." in *19th International Society for Music Information Retrieval Conference*. ISMIR, 2018, pp. 777–783.

[9] L. Henry, K. Frieler, G. Solis, M. Pfleiderer, S. Dixon, F. Höger, T. Weyde, and H.-C. Crayencour, "Dig that lick: Exploring patterns in jazz with computational methods," *Jazzforschung/Jazz Research, Vol. 50*, 2020.

[10] R. M. Bittner, J. Wilkins, H. Yip, and J. P. Bello, "MedleyDB 2.0: New data and a system for sustainable data collection," in *ISMIR Late Breaking and Demo Papers*. ISMIR, 2016.

[11] G. Peeters and K. Fort, "Towards a (better) definition of the description of annotated MIR corpora." in *13th International Society for Music Information Retrieval Conference*. ISMIR, 2012, pp. 25–30.

[12] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello, "Open-source practices for music signal processing research," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 128–137, 2018.

[13] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Creating DALI, a large dataset of synchronized audio, lyrics, and notes," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, 2020.

[14] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 1174–1178.

[15] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: An open-source library for sound and music analysis," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 855–858.

[16] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 659–663.

[17] E. J. Humphrey, J. Salamon, O. Nieto, J. Forsyth, R. M. Bittner, and J. P. Bello, "JAMS: A JSON annotated music specification for reproducible MIR research." in *15th International Society for Music Information Retrieval Conference*. ISMIR, 2014, pp. 591–596.

[18] A. Friberg and A. Sundström, "Swing ratios and ensemble timing in jazz performance," *Music Perception: An Interdisciplinary Journal*, vol. 19, no. 3, pp. 333–349, 2002.

[19] C. Dittmar, M. Pfleiderer, S. Balke, and M. Müller, "A swingogram representation for tracking micro-rhythmic variation in jazz performances," *Journal of New Music Research*, vol. 47, no. 2, pp. 97–113, 2018.

[20] C. Corcoran and K. Frieler, "Playing it straight: Analyzing jazz soloists' swing eighth-note distributions with the weimar jazz database," *Music Perception: An Interdisciplinary Journal*, vol. 38, no. 4, pp. 372–385, 2021.

[21] D. Stoller, S. Ewert, and S. Dixon, "Jointly detecting and separating singing voice: A multi-task approach," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 329–339.

[22] S. Balke, C. Dittmar, J. Abeßer, and M. Müller, "Data-driven solo voice enhancement for jazz music retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 196–200.

[23] J. Abeßer, K. Frieler, E. Cano, M. Pfleiderer, and W.-G. Zaddach, "Score-informed analysis of tuning, intonation, pitch modulation, and dynamics in jazz solos," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 168–177, 2017.

[24] R. M. Bittner, M. Fuentes, D. Rubinstein, A. Jansson, K. Choi, and T. Kell, "mirdata: Software for reproducible usage of datasets." in *19th International Society for Music Information Retrieval Conference*. ISMIR, 2019, pp. 99–106.