MedleyDB: A MULTITRACK DATASET FOR ANNOTATION-INTENSIVE MIR RESEARCH

Rachel Bittner[1], Justin Salamon[1,2], Mike Tierney[1], Matthias Mauch[3], Chris Cannam[3], Juan Bello[1]

[1]Music and Audio Research Lab, New York University

[2]Center for Urban Science and Progress, New York University

[3]Centre for Digital Music, Queen Mary University of London

{rachel.bittner,justin.salamon,mt2568,jpbello}@nyu.edu {m.mauch,chris.cannam}@eecs.qmul.ac.uk

ABSTRACT

We introduce *MedleyDB*: a dataset of annotated, royalty-free multitrack recordings. The dataset was primarily developed to support research on melody extraction, addressing important shortcomings of existing collections. For each song we provide melody f_0 annotations as well as instrument activations for evaluating automatic instrument recognition. The dataset is also useful for research on tasks that require access to the individual tracks of a song such as source separation and automatic mixing. In this paper we provide a detailed description of *MedleyDB*, including curation, annotation, and musical content. To gain insight into the new challenges presented by the dataset, we run a set of experiments using a state-of-the-art melody extraction algorithm and discuss the results. The dataset is shown to be considerably more challenging than the current test sets used in the MIREX evaluation campaign, thus opening new research avenues in melody extraction research.

1. INTRODUCTION

Music Information Retrieval (MIR) relies heavily on the availability of annotated datasets for training and evaluating algorithms. Despite efforts to crowd-source annotations [9], most annotated datasets available for MIR research are still the result of a manual annotation effort by a specific researcher or group. Consequently, the size of the datasets available for a particular MIR task is often directly related to the amount of effort involved in producing the annotations.

Some tasks, such as cover song identification or music recommendation, can leverage weak annotations such as basic song metadata, known relationships or listening patterns oftentimes compiled by large music services such as *last.fm* [1] . However, there is a subset of MIR tasks dealing with detailed information from the music signal for which time-aligned annotations are not readily available, such as the fundamental frequency (f_0) of the melody (needed for melody extraction [13]) or the activation times of the different instruments in the mix (needed for instrument recognition [1]). Annotating this kind of highly specific information from real world recordings is a time consuming process that requires qualified individuals, and is usually done in the context of large annotation efforts such as the Billboard [3], SALAMI [15], and Beatles [8] datasets. These sets include manual annotations of structure, chords, or notes, typically consisting of categorical labels at time intervals on the order of seconds. The annotation process is even more time-consuming for f_0 values or instrument activations for example, which are numeric instead of categorical, and at a time-scale on the order of milliseconds. Unsurprisingly, the datasets available for evaluating these taks are often limited in size (on the order of a couple dozen files) and comprised solely of short excerpts.

When multitrack audio is available, annotation tasks that would be difficult with mixed audio can often be expedited. For example, annotating the f_0 curve for a particular instrument from a full audio mix is difficult and tedious, whereas with multitrack stems the process can be partly automated using monophonic pitch tracking techniques. Since no algorithm provides 100% estimation accuracy in real-world conditions, a common solution is to have experts manually correct these machine annotations, a process significantly simpler than annotating from scratch. Unfortunately, collections of royalty-free multitrack recordings that can be shared for research purposes are relatively scarce, and those that exist are homogeneous in genre. This is a problem not only for evaluating annotation-intensive tasks but also for tasks that by definition require access to the individual tracks of a song such as source separation and automatic mixing.

In this paper we introduce *MedleyDB*: a multipurpose audio dataset of annotated, royalty-free multitrack recordings. The dataset includes melody f_0 annotations and was primarily developed to support research on melody extraction and to address important shortcomings of the existing collections for this task. Its applicability extends to research on other annotation-intensive MIR tasks, such as instrument recognition, for which we provide instrument activations. The dataset can also be directly used for re-

[1] http://www.last.fm

search on source separation and automatic mixing. Further track-level annotations (e.g. multiple f_0 or chords) can be easily added in the future to enable evaluation of additional MIR tasks.

The remainder of the paper is structured as follows: in Section 2 we provide a brief overview of existing datasets for melody extraction evaluation, including basic statistics and content. In Section 3 we provide a detailed description of the *MedleyDB* dataset, including compilation, annotation, and content statistics. In Section 4 we outline the types of annotations provided and the process by which they were generated. In Section 5 we provide some insight into the challenges presented by this new dataset by examining the results obtained by a state-of-the-art melody extraction algorithm. The conclusions of the paper are provided in Section 6.

2. PRIOR WORK

2.1 Datasets for melody extraction

Table 1 provides a summary of the datasets commonly used for the benchmarking of melody extraction algorithms. It can be observed that datasets that are stylistically varied and contain “real” music (e.g. ADC2004 and MIREX05) are very small in size, numbering no more than two dozen files and a few hundred seconds of audio. On the other hand, large datasets such as MIREX09, MIR1K and the RWC pop dataset tend to be stylistically homogeneous and/or include music that is less realistic. Furthermore, all datasets, with the exception of RWC, are limited to relatively short excerpts. Note that the main community evaluation for melody extraction, the MIREX AME task, [2] has been limited to the top 4 datasets.

In [14], the authors examined how the aforementioned constraints affect the evaluation of melody extraction algorithms. Three aspects were studied – inaccuracies in the annotations, the use of short excerpts instead of full-length songs, and the limited number of excerpts used. They found that the evaluation is highly sensitive to systematic annotation errors, that performance on excerpts is not necessarily a good predictor for performance on full songs, and that the collections used for the MIREX evaluation [5] are too small for the results to be statistically stable. Furthermore, they noted that the only MIREX dataset that is sufficiently large (MIREX 2009) is highly homogeneous (Chinese pop music) and thus does not represent the variety of commercial music that algorithms are expected to generalize to. This finding extrapolates to the MIR1K and RWC sets.

To facilitate meaningful future research on melody extraction, we sought to compile a new dataset addressing the following criteria:

1. **Size**: the dataset should be at least one order of magnitude greater than previous heterogeneous datasets such as ADC2004 and MIREX05.

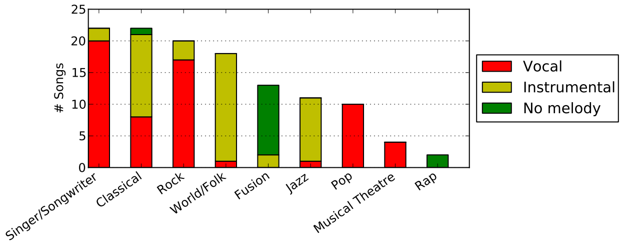

Figure 1. Number of songs per genre with breakdown by melody source type.

2. **Duration**: the dataset should primarily consist of full length songs.
3. **Quality**: the audio should be of professional or near-professional quality.
4. **Content**: the dataset should consist of songs from a variety of genres.
5. **Annotation**: the annotations must be accurate and well-documented.
6. **Audio**: each song and corresponding multitrack session must be available and distributable for research purposes.

2.2 Multitrack datasets

Since we opted to use multitracks to facilitate the annotation process, it is relevant to survey what multitrack datasets are currently available to the community. The TRIOS [6] dataset provides 5 score-aligned multitrack recordings of musical trios for source separation, the MASS [3] dataset contains a small collection of raw and effects-processed multitrack stems of musical excerpts for work in source separation, and the Mixploration dataset [4] for automatic mixing contains 24 versions of four songs. These sets are too small and homogeneous to fit our criteria; the closest candidate is the Structural Segmentation Multitrack Dataset [7] which contains 103 rock and pop songs with structural segmentation annotations. While the overall size of this dataset is satisfactory, there is little variety in genre and the dataset is not uniformly formatted, making batched processing difficult or impossible.

Since no sufficient multitrack dataset currently exists, we curated *MedleyDB* which fits our needs and can be used for other MIR tasks as well, and is described in detail in the following section.

3. DATASET

3.1 Overview

The dataset consists of 122 songs, 108 of which include melody annotations. The remaining 14 songs do not have a discernible melody and thus were not appropriate for melody extraction. We include these 14 songs in the dataset because of their use for other applications including instrument ID, source separation and automatic mixing.

[2] `http://www.music-ir.org/mirex/wiki/Audio_Melody_Extraction`

[3] `http://mtg.upf.edu/download/datasets/mass`

Name	# Songs	Song duration	Total duration	% Vocal Songs	Genres	Content
ADC2004	20	∼20 s	369 s	60%	Pop, jazz, opera	Real recordings, synthesized voice and MIDI
MIREX05	25	∼10–40 s	686 s	64%	Rock, R&B, pop, jazz, solo classical piano	Real recordings, synthesized MIDI
INDIAN08	8	∼60 s	501 s	100%	North Indian classical music	Real recordings
MIREX09	374	∼20–40 s	10020 s	100%	Chinese pop	Recorded singing with karaoke accompaniment
MIR1K	1000	∼10 s	7980 s	100%	Chinese Pop	Recorded singing with karaoke accompaniment
RWC	100	∼240 s	24403 s	100%	Japanese Pop, American Pop	Real recordings
MedleyDB	108	∼20–600 s	26831 s	57%	Rock, pop, classical, jazz, rock, pop, fusion, world, musical theater, singer-songwriter	Real recordings

Table 1. Existing collections for melody extraction evaluation (ADC2004 through RWC) and the new *MedleyDB* dataset.

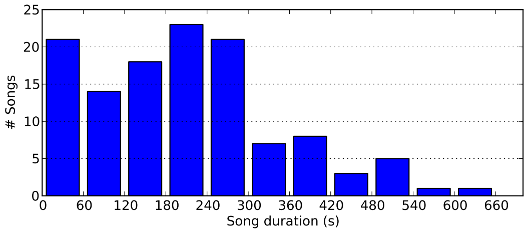

Figure 2. Distribution of song durations.

Each song in the dataset is freely available online[4] under a Creative Commons Attribution - NonCommercial - ShareAlike 3.0 Unported license[5], which allows the release of the audio and annotations for non-commercial purposes.

We provide a stereo mix and both dry and processed multitrack stems for each song. The content was obtained from multiple sources: 30 songs were provided by various independent artists, 32 were recorded at NYU's Dolan Recording Studio, 25 were recorded by Weathervane Music[6], and 35 were created by Music Delta[7]. The majority of the songs were recorded in professional studios and mixed by experienced engineers.

In Figure 1 we give the distribution of genres present within the dataset, as well as the number of vocal and instrumental songs within each genre. The genres are based on nine generic genre labels. Note that some genres such as Singer/Songwriter, Rock and Pop are strongly dominated by vocal songs, while others such as Jazz and World/Folk are mostly instrumental. Note that the Rap and most of the Fusion songs do not have melody annotations. Figure 2 depicts the distribution of song durations. A total of 105 out of the 122 songs in the dataset are full length songs, and the majority of these are between 3 and 5 minutes long. Most recordings that are under 1 minute long were created by Music Delta. Finally, the most represented instruments in the dataset are shown in Figure 3. Unsurprisingly, drums, bass, piano, vocals and guitars dominate the distribution.

3.2 Multitrack Audio Structure

The structure of the audio content in *MedleyDB* is largely determined by the recording process, and is exemplified in Figure 4, which gives a toy example of how the data could be organized for a recording of a jazz quartet.

At the lowest level of the process, a set of microphones is used to record the audio sources, such that there may be more than one microphone recording a single source – as is the case for the piano and drum set in Figure 4. The resulting files are *raw* unprocessed mono audio tracks. Note that while they are "unprocessed", they are edited such that there is no content present in the raw audio that is not used in the mix. The raw files are then grouped into stems, each corresponding to a specific sound source: double bass, piano, trumpet and drum set in the example. These *stems* are stereo audio components of the final mix and include all effects processing, gain control, and panning. Finally, we refer to the *mix* as the complete polyphonic audio created by mixing the stems and optionally mastering the mix.

Therefore, a song consists of the *mix*, *stems*, and *raw audio*. This hierarchy does not perfectly model every style of recording and mixing, but it works well for the majority of songs. Thus, the audio provided for this dataset is organized with this hierarchy in mind.

3.3 Metadata

Both song and stem-level metadata is provided for each song. The song-level metadata includes basic information about the song such as the artist, title, composer, and website. Additionally, we provide genre labels corresponding to the labels in Figure 1. Some sessions correspond to

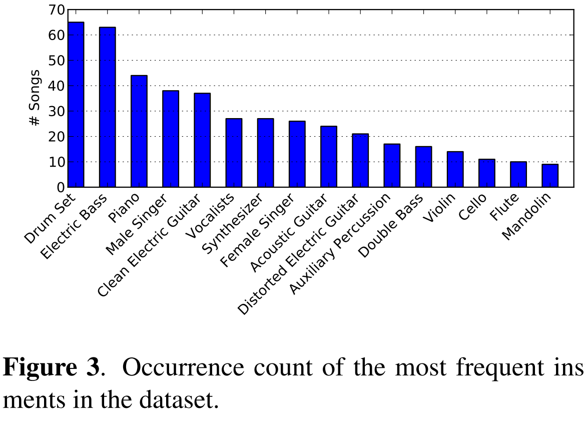

Figure 3. Occurrence count of the most frequent instruments in the dataset.

[4] http://marl.smusic.nyu.edu/medleydb

[5] http://creativecommons.org/licenses/by-nc-sa/3.0/deed.en_US

[6] http://weathervanemusic.org/

[7] http://www.musicdelta.com/

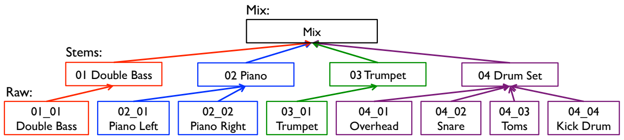

Figure 4. The hierarchy of audio files for a jazz quartet.

recordings of ensembles, where the microphones may pick up sound from sources other than the one intended, a phenomenon known as *bleeding*. Because bleed can affect automated annotation methods and other types of processing, songs that contain any stems with bleed are tagged.

Stem-level metadata includes instrument labels based on a predefined taxonomy given to annotators, and a field indicating whether the stem contains melody.

The metadata is provided as a YAML[8] file, which is both human-readable as a text file, and a structured format that can be easily loaded into various programming environments.

4. ANNOTATIONS

4.1 Annotation Task Definitions

When creating annotations for *MedleyDB*, we were faced with the question of what definition of melody to use. The definition of melody used in MIREX 2014 defines melody as the predominant pitch where, "pitch is expressed as the fundamental frequency of the main melodic voice, and is reported in a frame-based manner on an evenly-spaced time-grid." Many of the songs in the dataset do not reasonably fit the definition of melody used by MIREX because of the constraint that the melody is played by a single voice, but we felt that the annotations should have consistency with the existing melody annotations.

Our resolution was to provide melody annotations based on three different definitions of melody that are in discussion within the MIR community.[9] In the definitions we consider, melody is defined as:

1. The f_0 curve of the predominant melodic line drawn from a single source.
2. The f_0 curve of the predominant melodic line drawn from multiple sources.
3. The f_0 curves of all melodic lines drawn from multiple sources.

Definition 1 coincides with the definition for the melody annotations used in MIREX. This definition requires the choice of a lead instrument and gives the f_0 curve for this instrument. Definition 2 expands on definition 1 by allowing multiple instruments to contribute to the melody. While a single lead instrument need not be chosen, an indication of which instrument is predominant at each point in time is required to resolve the f_0 curve to a single point at each time frame. Definition 3 is the most complex, but also the most general. The key difference in this definition is that at a given time frame, multiple f_0 values may be "correct".

For instrument activations, we simply assume that signal energy in a given stem, above a predefined limit, is indicative of the presence of the corresponding instrument in the mix. Based on this notion, we provide two types of annotations: a list of time segments where each instrument is active; and a matrix containing the activation confidence per instrument per unit of time.

4.2 Automatic Annotation Process

The melody annotation process was semi-automated by using monophonic pitch tracking on selected stems to return a good initial estimate of the f_0 curve, and by using a voicing detection algorithm to compute instrument activations. The monophonic pitch tracking algorithm used was pYIN [11] which is an improved, probabilistic version of the well-known YIN algorithm.

As discussed in the previous section, for each song we provide melody annotations based upon the 3 different definitions. The melody annotations based on Definition 1 were generated by choosing the single most dominant melodic stem. The Definition 2 annotations were created by sectioning the mix into regions and indicating the predominant melodic stem within each region. The melody curve was generated by choosing the f_0 curve from the indicated instrument at each point in time. The Definition 3 annotations contain the f_0 curves from each of the annotated stems.

The annotations of instrument activations were generated using a standard envelope following technique on each stem, consisting of half-wave rectification, compression, smoothing and down-sampling. The resulting envelopes are normalized to account for overall signal energy and total number of sources, resulting in the $t \times m$ matrix H, where t is the number of analysis frames, and m is the number of instruments in the mix. For the i^{th} instrument, the confidence of its activations as a function of time can be approximated via a logistic function:

$$C(i,t) = 1 - \frac{1}{1 + e^{(H_{it} - \theta)\lambda}}. \tag{1}$$

where λ controls the slope of the function, and θ the threshold of activation. Frames where instrument i is considered active are those for which $C(i,t) \geq 0.5$. No manual correction was performed on these activations.

Note that monophonic pitch tracking, and the automatic detection of voicing and instrument activations, fail when the stems contain bleed from other instruments, which is the case for 25 songs within the collection. Source separation, using a simple approach based on Wiener filters [2], was used on stems with bleed to clean up the audio before applying the algorithms. The parameters of the separation were manually and independently optimized for each track containing bleed.

[8] http://www.yaml.org/

[9] http://ameannotationinitiative.wikispaces.com

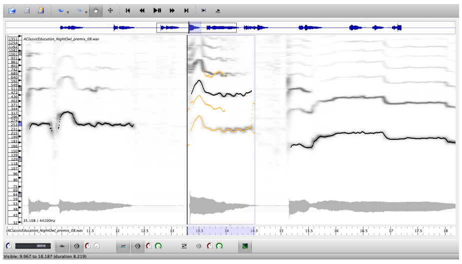

Figure 5. Screenshot of *Tony*. An estimated pitch curve is selected and alternative candidates are shown in yellow.

4.3 Manual Annotation Process

The manual annotation process was facilitated by the use of a recently developed tool called *Tony* [10], which enables efficient manual corrections (see Figure 5). Tony provides 3 types of semi-manual correction methods: (1) deletion (2) octave shifting and (3) alternative candidates.

When annotating the f_0 curves, unvoiced vocal sounds, percussive attacks, and reverb tail were removed. Sections of a stem which were active but did not contain melody were also removed. For example, a piano stem in a jazz combo may play the melody during a solo section and play background chords throughout the rest of the piece. In this case, only the solo section would be annotated, and all other frames would be marked as unvoiced.

The annotations were created by five annotators, all of which were musicians and had at least a bachelor's degree in music. Each annotation was evaluated by one annotator and validated by another. The annotator/validator pairs were randomized to make the final annotations as unbiased as possible.

4.4 Annotation Formats

We provide melody annotations based on the three definitions for 108 out of the 122 songs. Note that while definition 1 is not appropriate for all of the annotated songs (i.e. there are songs where the melody is played by several sources and there is no single clear predominant source throughout the piece), we provide type 1 melody annotations for all 108 melodic tracks so that an algorithm's performance on type 1 versus type 2 melody annotations can be compared over the full dataset. Of the 108 songs with melody annotations, 62 contain predominantly vocal melodies and the remaining 47 contain instrumental melodies.

Every melody annotation begins at time 0 and has a hop size of 5.8 ms (256 samples at $f_s = 44.1$ kHz). Each time stamp in the annotation corresponds to the center of the analysis frame (i.e. the first frame is centered on time 0). In accordance with previous annotations, frequency values are given in Hz, where unvoiced frames (i.e. frames where there is no melody) are indicated by a value of 0 Hz.

We provide instrument activation annotations for the entire dataset. Confidence values are given as matrices where the first column corresponds to time in seconds, starting at 0 with a hop size of 46.4 ms (2048 samples at $f_s = 44.1$ kHz), and each subsequent column corresponds to an instrument identifier. Confidence values are continuous in the range $[0, 1]$. We also provide a list of activations, each a triplet of start time, end time and instrument label.

Dataset	ν	**VxR**	**VxF**	**RPA**	**RCA**	**OA**
MDB – All	.2	.78 (.13)	.38 (.14)	.55 (.26)	.68 (.19)	**.54** (.17)
MDB – All	-1	.57 (.20)	.20 (.12)	.52 (.26)	.68 (.19)	**.57** (.18)
MDB – VOC	-1	.69 (.15)	.23 (.13)	.63 (.23)	.76 (.15)	**.66** (.14)
MDB – INS	-1	.41 (.15)	.16 (.09)	.38 (.23)	.57 (.18)	**.47** (.17)
MIREX11	.2	.86	.24	.80	.82	**.75**

Table 2. Performance of Melodia [12] on different subsets of *MedleyDB* (*MDB*) for type 1 melody annotations, and comparison to performance on the MIREX datasets. For each measure we provide the mean with the standard deviation in parentheses.

5. NEW CHALLENGES

To gain insight into the challenges presented by this new dataset and its potential for supporting progress in melody extraction research, we evaluate the performance of the Melodia melody extraction algorithm [12] on the subset of *MedleyDB* containing melody annotations. In the following experiments we use the melody annotations based on Definition 1, which can be evaluated using the standard five measures used in melody extraction evaluation: voicing recall (VxR), voicing false alarm (VxF), raw pitch accuracy (RPA), raw chroma accuracy (RCA), and overall accuracy (OA). For further details about the measures see [13].

In the first row of Table 2 we give the results obtained by Melodia using the same parameters (voicing threshold $\nu = .2$) employed in MIREX 2011 [12]. The first thing we note is that for all measures, the performance is considerably lower on *MedleyDB* than on MIREX11. The overall accuracy is 21 percentage points lower, a first indication that the new dataset is more challenging. We also note that the VxF rate is considerably higher compared to the MIREX results. In the second row of Table 2 we provide the results obtained when setting ν to maximize the overall accuracy ($\nu = -1$). The increase in overall accuracy is relatively small (3 points), indicating that the dataset remains challenging despite using the best possible voicing parameter. In the next two rows of Table 2, we provide a breakdown of the results by vocal vs. instrumental songs. We see that the algorithm does significantly better on vocal melodies compared to instrumental ones, consistent with the observations made in [12]. For instrumental melodies we observe a 19-point drop between raw chroma and pitch accuracy, indicating an increased number of octave errors. The bias in performance towards vocal melodies is likely the result of all previous datasets being primarily vocal.

In Table 3 we provide a breakdown of the results by genre. In accordance with the the previous table, we see that genres with primarily instrumental melodies are considerably more challenging. Finally, we repeat the experiment carried out in [14], where the authors compared performance on recordings to shorter sub-clips taken from the same recordings to see whether the results on a dataset of

Genre	VxR	VxF	RPA	RCA	OA
MUS	.73 (.16)	.14 (.04)	.74 (.18)	.87 (.08)	**.73** (.14)
POP	.74 (.12)	.22 (.09)	.65 (.20)	.73 (.15)	**.69** (.12)
S/S	.66 (.13)	.23 (.12)	.64 (.19)	.74 (.16)	**.66** (.11)
ROC	.71 (.18)	.29 (.15)	.53 (.29)	.73 (.18)	**.59** (.16)
JAZ	.44 (.14)	.12 (.06)	.55 (.17)	.68 (.15)	**.57** (.14)
CLA	.46 (.20)	.15 (.07)	.35 (.30)	.56 (.22)	**.51** (.23)
WOR	.40 (.12)	.18 (.09)	.44 (.19)	.63 (.14)	**.44** (.13)
FUS	.41 (.04)	.17 (.02)	.32 (.07)	.51 (.01)	**.43** (.04)

Table 3. Performance of Melodia [12] ($\nu = -1$) on different genres in *MedleyDB* for type 1 melody annotations. For each measure we provide the mean with the standard deviation in parentheses.

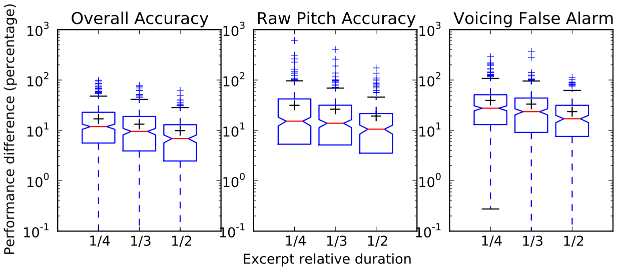

Figure 6. Relative performance differences between full songs and excerpts. The large black crosses mark the means of the distributions.

excerpts would generalize to a dataset of full songs. The novelty in our experiment is that we use full length songs, as opposed to clips sliced into even shorter sub-clips. The results are presented in Figure 6, and are consistent with those reported in [14]. We see that as the relative duration of the excerpts (1/4, 1/3 or 1/2 of the full song) gets closer to 1, the relative difference in performance goes down (significant by a Mann-Whitney U test, $\alpha = 0.01$). This highlights another benefit of *MedleyDB*: since the dataset primarily contains full length songs, one can expect better generalization to real-world music collections. While further error analysis is required to understand the specific challenges presented by *MedleyDB*, we identify (by inspection) some of the musical characteristics across the dataset that make *MedleyDB* more challenging – rapidly changing notes, a large melodic frequency range (43-3662 Hz), concurrent melodic lines, and complex polyphony.

6. CONCLUSION

Due to the scarcity of multitrack audio data for MIR research, we presented *MedleyDB* – a dataset of over 100 multitrack recordings of songs with melody f_0 annotations and instrument activations. We provided a description of the dataset, including how it was curated, annotated, and its musical content. Finally, we ran a set of experiments to identify some of the new challenges presented by the dataset. We noted how the increased proportion of instrumental tracks makes the dataset significantly more challenging compared to the MIREX datasets, and confirmed that performance on excerpts will not necessarily generalize well to full-length songs, highlighting the greater generalizability of *MedleyDB* compared with most existing datasets. Since 2011 there has been no significant improvement in performance on the MIREX AME task. If we previously attributed this to some glass ceiling, we now see that there is still much room for improvement. *MedleyDB* represents a shift towards more realistic datasets for MIR research, and we believe it will help identify future research avenues and enable further progress in melody extraction research and other annotation-intensive MIR endeavors.

7. REFERENCES

[1] J.G.A. Barbedo. Instrument recognition. In T. Li, M. Ogihara, and G. Tzanetakis, editors, *Music Data Mining*. CRC Press, 2012.

[2] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE TASLP*, 14(1):191–199, 2006.

[3] J. A. Burgoyne, J. Wild, and I. Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *ISMIR'11*, pages 633–638, 2011.

[4] M. Cartwright, B. Pardo, and J. Reiss. Mixploration: Rethinking the audio mixer interface. In *19th Int. Conf. on Intelligent User Interfaces*, pages 365–370, 2014.

[5] J. Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[6] J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *IEEE ICASSP'13*, pages 888–891, 2013.

[7] S. Hargreaves, A. Klapuri, and M. Sandler. Structural segmentation of multitrack audio. *IEEE TASLP*, 20(10):2637–2647, 2012.

[8] C. Harte, M. B. Sandler, S. A Abdallah, and E. Gómez. Symbolic representation of musical chords: A proposed syntax for text annotations. In *ISMIR'05*, pages 66–71, 2005.

[9] M. I. Mandel and D. P. W. Ellis. A web-based game for collecting music metadata. *J. of New Music Research*, 37(2):151–165, 2008.

[10] M. Mauch and G. Cannam, C. Fazekas. Efficient computer-aided pitchtrack and note estimation for scientific applications, 2014. SEMPRE'14, extended abstract.

[11] M. Mauch and S. Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE ICASSP'14*, 2014. In press.

[12] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE TASLP*, 20(6):1759–1770, 2012.

[13] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard. Melody extraction from polyphonic music signals: Approaches, applications and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134, 2014.

[14] J. Salamon and J. Urbano. Current challenges in the evaluation of predominant melody extraction algorithms. In *ISMIR'12*, pages 289–294, 2012.

[15] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J .S. Downie. Design and creation of a large-scale database of structural annotations. In *ISMIR'11*, pages 555–560, 2011.