ENSEMBLE: A HYBRID HUMAN-MACHINE SYSTEM FOR GENERATING MELODY SCORES FROM AUDIO

Tim Tse[1], Justin Salamon[2], Alex Williams[1], Helga Jiang[1] and Edith Law[1]
[1]University of Waterloo, [2]New York University
{trttse, alex.williams, hhjiang, edith.law}@uwaterloo.ca, justin.salamon@nyu.edu

ABSTRACT

Music transcription is a highly complex task that is difficult for automated algorithms, and equally challenging to people, even those with many years of musical training. Furthermore, there is a shortage of high-quality datasets for training automated transcription algorithms. In this research, we explore a *semi-automated, crowdsourced* approach to generate music transcriptions, by first running an automatic melody transcription algorithm on a (polyphonic) song to produce a series of discrete notes representing the melody, and then soliciting the crowd to correct this melody. We present a novel web-based interface that enables the crowd to correct transcriptions, report results from an experiment to understand the capabilities of non-experts to perform this challenging task, and characterize the characteristics and actions of workers and how they correlate with transcription performance.

1. INTRODUCTION

Music transcription is the process of transforming the acoustic representation of a music piece to a notational representation (e.g., music score). Despite active research on automating this process, music transcription remains a difficult problem [1, 13] for automated algorithms, and equally challenging for human annotators, even those with formal music training. As a result, there is a lack of scalable methods for generating ground truth datasets for training and evaluating music transcription algorithms.

Crowdsourcing has demonstrated great promise as an avenue for generating large datasets. Recent work suggests that the crowd may be capable of performing tasks that require expert knowledge [24, 30]. In this paper, we investigate whether it is feasible to elicit the help of the non-expert crowd to streamline the generation of music transcription ground truth data. Specifically, we introduce a *semi-automated* system, which first extracts a note representation of the melody from an audio file, and then solicits the crowd to correct the melody by making small adjustments to the pitch, onset and offset of each note. Our goal is to characterize the extent to which the crowd can successfully perform this complex task which typically requires musical expertise, describe the relationship between the actions that users take to correct transcriptions and their relationship to performance, and based on these findings, highlight specific challenges associated with crowdsourcing music transcription.

2. RELATED WORK

There are three threads of prior work relevant to our research: automatic music transcription (AMT), semi-automatic music transcription, and crowdsourcing the generation of ground truth data in music information retrieval.

In this work, we tackle a specific subtask of automatic music transcription, automatic *melody* transcription. Given a polyphonic music recording that has a clear melodic line, our goal is to transcribe the melody into a piano-roll like representation consisting of a sequence of non-overlapping notes, each with an onset time, offset time, and a pitch value. While several approaches have been proposed to date (see [27] and references therein), the task remains highly challenging and is considered an open problem.

Given that fully automated transcription techniques are at the moment still limited, it is worth investigating how well machines aided by humans, or *semi-automatic* systems, perform at this task. In [15] the authors studied two types of user input for semi-automatic music transcription based on matrix deconvolution. They showed that by asking the users to transcribe a small number of notes from the test data, performance could be significantly improved compared to initializing the model from independent training data. [8] introduced a human-in-the-loop model which solicits users to highlight notes to be extracted from the rest of the audio. In [29], users are asked to hum the melody to be extracted in a sound mixture. A source separation framework that incorporates prior knowledge has been proposed by [20], and the authors have shown that the informed settings outperform the blind settings. Finally, recent work [16] elicits help from 30 users to provide note onsets and pitches as seeds to a semi-automated melody extraction algorithm, and found that experts and novices alike were able to contribute useful information. Likewise, Songle [12], a web service for music listening, provides users with the ability to modify a draft transcription generated by an automated algorithm.

There have been a number of crowdsourcing experiments in music information retrieval that have investigated how to generate ground truth datasets for music tagging [17, 32] and music similarity judgments [33], and for more complicated music information retrieval tasks, such as the creation of emotionally relevant musical scores for audio stories [25]. These prior works found that the non-expert crowd can be incentivized to contribute accurate annotations for music.

3. ENSEMBLE

We propose Ensemble, a two-part architecture to the task of music transcription (Figure 1). The first part involves the task of automatically transcribing the notes of the melody from the audio signal of a music piece. The second part asks the crowd to fix and improve upon the automatically generated score.

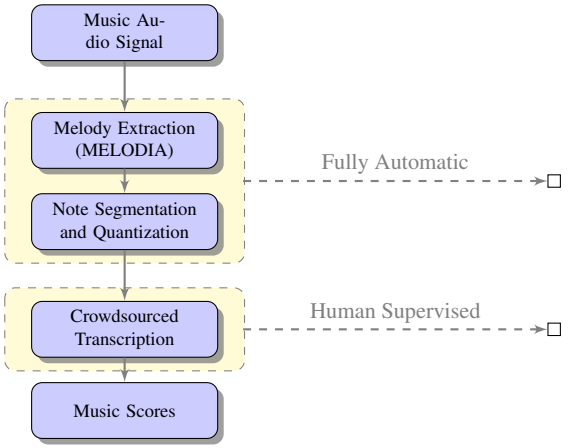

Figure 1. A semi-automatic architecture. Melody extraction and quantization is performed automatically, and this automatically generated transcription is then corrected by the crowd.

3.1 Automatic Melody Note Transcription

We employ a two stage process to produce an automatic transcription of the melody notes from the audio signal. First, we extract the continuous fundamental frequency (f_0) contour of the melody using the MELODIA melody extraction plugin [28]. Next, we quantize the contour in pitch and segment it in time to produce a set of discrete notes. This is performed by means of the following set of simple heuristics: first, every f_0 value is mapped to its nearest semitone pitch assuming an equally tempered scale tuned to 440 Hz. The sequence is then smoothed using a median filter of 250 ms to avoid very short pitch jumps that can occur due to e.g. the presence of vibrato in the unquantized pitch contour. Finally, the sequence must be segmented in time into notes. Since MELODIA already estimates the start and end times of each voiced section (i.e. sections where the melody is present), we only need to identify note transitions within each section, accomplished by simply starting a new note whenever the quantized pitch value changes from one semitone to another. We impose a minimum note duration of 100 ms to avoid very short notes generated by continuous pitch transitions such as glissando. It should be noted that more advanced note segmentation algorithms have been proposed [10, 19], but since our goal is to evaluate the capability of the non-expert crowd to correct the automatic transcriptions (not solve automatic note segmentation), we do not require a state-of-the-art method for this study. Our complete melody note transcription script is available online [1].

3.1.1 Evaluation Metrics and Tools

To evaluate the agreement between two note transcriptions (i.e., automated/crowdsourced transcription against ground truth), we use the evaluation metrics from the MIREX [5] Note Tracking subtask of the Multiple Fundamental Frequency Estimation & Tracking challenge [2]:

$$\text{Precision} = \frac{|\text{correct estimated notes}|}{|\text{estimated notes}|} \quad (1)$$

$$\text{Recall} = \frac{|\text{correct estimated notes}|}{|\text{reference notes}|} \quad (2)$$

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

An estimated note is considered to match a reference (ground truth) note if its pitch is within half a semitone (± 50 cents) of the reference pitch, its onset is within ± 50 ms of the reference note's onset, and its offset is within ± 50 ms or 20% of the reference note's duration from the reference note's offset, whichever is greater. MIREX also computes a second version of each metric where note offsets are ignored for note matching, since offsets are both considerably harder to detect automatically and more subjective: our ability to perceive offsets can be strongly affected by the duration of note decay, reverberation and masking [3, 18]. In light of this, and following our own difficulty in annotating offsets for our dataset, for this study we use the metrics that ignore note offsets, as we do not consider it reasonable to expect our participants to be able to accurately match the offsets when our expert was not certain about them in the first place. The metrics are computed using the JAMS [14] evaluation wrapper for the `mir_eval` library [22].

3.2 Crowdsourcing Interface

Figure 2 depicts the interface that we designed for crowd workers to correct the music transcription. The interface consists of two panels: the reference panel (top) and the transcription panel (bottom). The reference panel displays the waveform of the original music clip. The transcription panel displays the current note transcription; initially, this is the transcription automatically generated by the

[1] `https://github.com/justinsalamon/audio_to_midi_melodia`

[2] `http://www.music-ir.org/mirex/wiki/2015:Multiple_Fundamental_Frequency_Estimation_\%26_Tracking_Results_-_MIREX_Dataset#Task_2:Note_Tracking_.28NT.29`

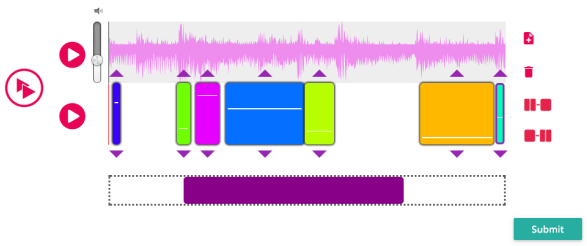

Figure 2. Screenshot of the note editing interface.

MELODIA-based transcription algorithm. Similar to many music editing software, the transcription panel uses rectangles to denote individual notes, height of the white bar within the rectangle (and in our case, also rectangle color) to denote pitch height, and width to denote duration. Workers are given controls to play the reference music clip and transcription separately, as well as simultaneously.

To edit the current transcription, workers can add, delete, split and merge notes. Notes can be added either by clicking on the "add note" button on the right panel or by double-clicking the area where the user wishes to add a note. Notes can be deleted by first clicking on a note, then either clicking the "delete note" button or tapping "backspace" on the keyboard. Adjacent notes can be merged together by clicking on the first note (in time) and then clicking on the "merge notes" button. Similarly, a note can be split into two notes by first clicking on it and then clicking the "split note" button. Pitch can be adjusted (by one semitone at a time) either by using the up and down arrows attached to each note, or by clicking on the note and then pressing the up and down arrow keys on the keyboard. Every time the pitch is adjusted the note is played back to indicate the new pitch. The note can be moved in time by dragging it left or right, and the onset and offset can be adjusted by dragging the note edges left or right. The purple bar at the bottom determines which segment of the audio/transcription is to be played. The bar is adjustable hence allowing the user to selectively choose the portion of the audio to focus on and compare the corrected transcription against.

4. STUDY DESIGN

In this work, our goal is to understand the extent to which crowdworkers can perform transcription correction tasks, which typically requires expertise. To do this, we conducted an experiment via Amazon Mechanical Turk to engage crowdworkers to correct the automatically transcribed melody from a 30 second excerpt of a song. Each 30 s excerpt is sliced into ten 3 s music clips, and each Turker performed corrections on all ten clips, one clip at a time. The reason we use short 3 s clip is that the results of a previous in-lab pilot test suggested that longer clips (10 s) resulted in too much cognitive load on the worker, rendering the task too overwhelming for them. Furthermore, Mechanical Turk workers are more used to performing a series of short, *micro*-tasks.

4.1 Data

For our experiments, we require a dataset of polyphonic music that has clear melodies (such as popular music) with ground truth annotations on a note level. Surprisingly, we found it hard to find suitable data. Most datasets with note annotations are comprised of music that does not contain an unambiguous melody (e.g. chamber music [7, 31] or solo piano [6]), or contain artificially synthesized audio data [9]. The two most relevant datasets, RockCorpus [4] and RWC-Pop [11] also turned out to be problematic: the former only contains pitch-class annotations for notes while the latter was problematic in terms of aligning the audio content to the provided MIDI annotations, which we found to be insufficiently accurate for the purpose of this study. Ultimately, we decided to create our own dataset.

To create the dataset, we selected 20 songs from the Million Song Dataset [2] which had a clear melody, in all cases sung by the human voice. For each song we obtained a MIDI file that was automatically aligned to a 30 s excerpt of the audio from [23]. We then manually identified the MIDI track containing the melody notes and separated it from the rest of the tracks using `pretty_midi` [21]. The separated melody was then loaded into Logic Pro and manually corrected by one of the authors with formal music training to match the melody as accurately as possible, and finally converted to JAMS format [14]. Since transcribing a sung melody into a series of quantized notes is a task that contains a certain degree of subjectivity, the following guidelines were followed:

- The onset and offset of each annotated note should match the audio recording as accurately as possible.
- Pitch is annotated on an equally-tempered semitone scale tuned to 440 Hz.
- Every sung syllable is annotated as a separate note.
- Whenever the pitch of the melody is perceived to change by a semitone or more it should be annotated as a separate note, even if the syllable does not change, including embellishments and glissandos.

The same guidelines were communicated to crowdworkers in the tutorial preceding the experiment.

4.2 Participants

There were a total of 105 Turkers who participated in the study. Each Turker was assigned a randomly chosen 30 s excerpt (sliced into 10 clips 3 s each). The majority of the workers were from the United States. The entire task takes roughly 20-30 minutes, and workers were paid $5 in total. Since we wish to evaluate how well a layperson randomly drawn from the crowd can perform this task, we did not restrict the pool of participants by the level of their formal music training. Finally, our system ensures that each worker takes our study only once.

4.3 Procedure

To begin, workers watch a tutorial video that shows how an expert corrects the transcription for a 3 s music clip. The tutorial guides new users on how to approach this task, by highlighting all of the functionalities of the interface and specifying the rules to follow when considering the correctness of their transcription. The tutorial is followed by a pre-study questionnaire, which asks workers about their music background, including the number of years of formal music training, a set of listening tests (i.e., listening to two consecutive notes and determining which one has a higher pitch), and a set of knowledge questions related to music reading (e.g., whether they are familiar with musical notation such as key and time signature). Workers performed a series of 10 tasks using the note editor interface to correct the transcription of the melody of consecutive 3 s music clips. In order to better understand worker behavior and intent, all interactions with the interface were recorded and saved in a database alongside the workers' corrected melody transcription.

After finishing the transcription, workers are asked to complete a post-study questionnaire, which asks them about their experience using our interface. In particular, we capture motivational factors using the Intrinsic Motivation Inventory (IMI) [26], a scale that measures factors related to *enjoyment* (how much workers enjoy the task), *competence* (how competent workers think they are at the task) and *effort* (how much effort workers put into the task). Workers are asked to rate how much they agree with a set of statements related to these three dimensions on a 7-point Likert scale. For each dimension, we then average the workers' responses (inverting the scores for the negative statements) and use the mean value as the summary statistic for that dimension. Finally, we ask workers to rate the difficulty of the task, as well as comment on ways in which the interface could be improved and the aspects of the task they found most challenging.

5. RESULTS

5.1 Worker Performance

To understand whether the crowd can improve the automatic melody transcription, for each 3 s music clip we compute the F-measure (ignoring offsets) of the automatic (AMT) and the crowd-generated transcriptions against the ground truth. In Figure 3(a) we present the scores of the crowd-annotated transcriptions (green dots) against those of the AMT (blue line) for each 3 s clip, ordered by the score obtained by the AMT. In Figure 3(b) we present the same results, but average the scores for each clip over all workers who annotated that clip. Note that if a data point is located above the $y = x$ line it means the crowd-annotated transcription outperformed the AMT, and vice versa.

The number of points above the line, on the line, and below the line are 272, 366, 338 respectively for subplot (a) and 85, 13, 94 respectively for subplot (b). In general we see that the worker scores vary a lot, with some workers able to correct the AMT to perfectly match the ground truth and others capable of "ruining" an already good AMT. The large number of points on the line in (a) suggests that oftentimes the workers did not think they can improve the AMT and left the transcription unchanged (or modified it insignificantly). In both subplots we observe that when the AMT score exceeds 0.6, the majority of the points fall below the line, suggesting that the crowd has a hard time improving the transcription for clips for which the AMT already performs relatively well. Conversely, when the AMT performs poorly the crowd is capable of making corrections that improve the transcription (i.e., the majority of the data points are on or above the line).

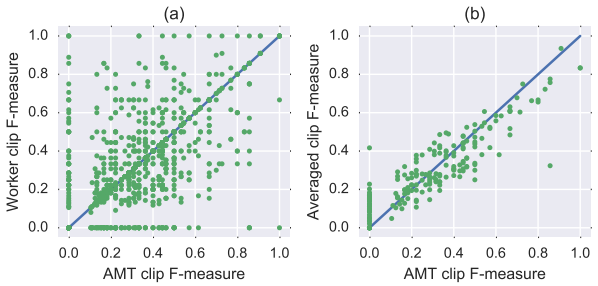

Figure 3. Worker clip F-measure against AMT clip F-measure: (a) individual worker-clip scores, (b) averaged per-clip scores.

5.2 Worker Characteristics

The answers to the pre-study questionnaire are summarized Figure 4: (a) shows the distribution of the workers' musical expertise (T_1), (b) the number of pitch comparison questions answered correctly (T_2), and (c) the number of music notation questions answered correctly (T_3).

We see that the majority of the workers have little to no formal music training with 62% responding "None" or "1 year". 93% of workers answered at least two pitch comparison correctly and 63% answered at least two musical knowledge questions correctly. Given the variability in the scores achieved by the workers, we wanted to see if there was correlation between the workers' answers and their F-measure performance. To determine this, in Figure 4 (d), (e), and (f) we plot the workers' F-measure performance against the three separate topics T_1, T_2 and T_3 respectively. We also compute their Pearson correlation coefficients: 0.12, 0.11 and 0.18 respectively. Results show that the factor most correlated to the workers' performance is their understanding of musical notation. A possible explanation is that the person's proficiency with musical notation is a good indicator of their actual musical expertise. Self-reported expertise is not as good an indicator: this could be (for example) because the worker's musical training happened a long time ago and has since deteriorated through disuse (e.g., an adult took formal music lessons when they were a child but never again). Interestingly, the ability to compare pitches also has a relatively low correlation to the F-measure performance. A possible explanation for this is that the comparison questions (determin-

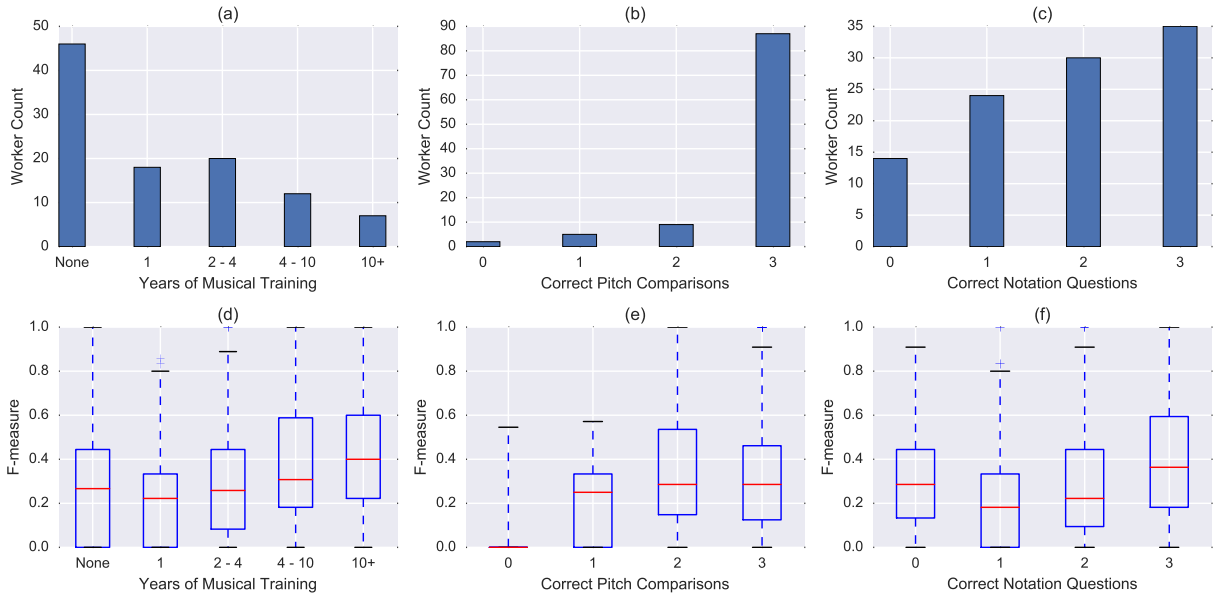

Figure 4. Pre-questionnaire results: (a) years of formal music training, (b) correctly answered pitch comparison questions, (c) correctly answered musical notation questions, (d) years of training vs. performance, (e) correctly answered pitch comparison questions vs. performance, (f) correctly answered musical notation questions vs. performance.

ing which of two pitches is higher) are easier than the task of matching the pitch of a synthesized note to the human voice.

The post-questionnaire shows that there are three main types of challenges. First, many workers ($\sim 20\%$) reported having difficulty matching the pitch of the transcription to the pitch in the audio (e.g., "It's hard to exactly match the notes to the melody."). There were several reasons cited, including *personal inexperience* (e.g., "I'm pretty sure I'm completely tone deaf. It was like trying to learn a foreign language."), *difficulty in separating the vocals from the background* (e.g., "Sometimes it was hard to hear the exact melody over the instrumentation.", "In my specific audio clips, the voice was very ambient and seemed to be layered in octaves at times. So, using only one tone to accurately denote pitch was slightly confusing.") and *difficulty with decorative notes* (e.g., "Vocal inflections/trails can be almost impossible to properly transcribe", "Getting all the nuances of the notes correctly, especially the beginning and ends of notes where the singer sort of "slides" into and out of the note.", "Some changes in tone were particularly difficult to find the right tone to match the voice. Mostly the 'guttural' parts of the singing.", "When someone's voice does that little vibrato tremble thing, it's almost impossible to accurately tab that out.").

Second, some workers reported difficulty with timing and rhythm, knowing *exactly* when things should start and end. For example, one musically trained worker said "Timing was incredibly difficult. I've been a (self taught) musician for 7 years so recognizing pitch wasn't difficult. The hard part was fitting the notes to the parts of the song in the right timing. If there's anything I messed up on during my work, it was that." Finally, a few workers mentioned finding the task difficult due to the lack of feedback indicating whether they were doing the task accurately ("It was hard to tell for sure if I was doing well or not.").

IMI Factor	Mean	Standard Deviation
Enjoyment	5.87	1.46
Competence	4.15	1.79
Effort	6.30	0.93

Table 1. Mean and standard deviation for each IMI factor.

Many workers describe the interface as "intuitive," "easy to work", "well made," and "straightforward". The most requested functionality was the ability to play the audio at a slower speed, e.g., so that one can "catch grace notes or time beats more precisely." In practice this would require us to incorporate a time-stretching algorithm into the interface. Other requests include the ability to duplicate a note, undo previous actions, and get more training and feedback.

Overall, there seems to be an interesting tension: workers find the task extremely challenging, yet enjoyable. For example, workers said "This is one of the best hits on Amazon mTurk.", "Getting it all to sound great is a challenge but fun", "Maybe I'm just not musically inclined, but even after going through several notes, it was still difficult for me to figure out whether or not they matched the singer's voice. Very challenging, but interesting too!"

The quantitative data also reflect this observation. Table 1 summarizes the average intrinsic motivational factors—enjoyment, competence and effort—over all workers. Results show that on average, workers enjoy the task (μ=5.87, σ=1.46), but at the same time, found themselves lacking competence (μ=4.15, σ=1.79) in this task that they find effortful (μ=6.30, σ=0.93). In addition, they reported finding the task difficult (μ=5.57, σ=1.55).

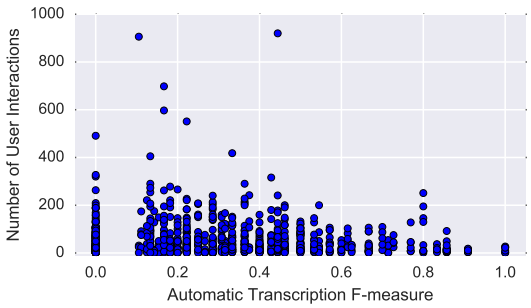

Figure 5. User interaction count vs. AMT performance.

Action	Occur.	Action	Occur.
Play Both	16.1 (26.2)	Resize Play Region	6.3 (8.0)
Change Pitch	12.7 (12.1)	Move Play Region	3.9 (3.8)
Change Offset	9.6 (10.6)	Add Note	2.3 (1.9)
Play Wav	8.3 (13.2)	Merge Note	2.1 (1.8)
Change Onset	7.5 (9.0)	Delete Note	2.0 (1.4)
Play Transcription	7.2 (10.5)	Split Note	1.8 (1.1)

Table 2. Average occurrence (and standard deviation) of each worker action for all 3-second clips.

5.3 Worker Actions

In order to understand more deeply how workers perform the transcriptions, we analyze the action log consisting of a list of coded user actions, including changes in onset, offset, or pitch (up or down), note addition, deletion, merge and split, as well as actions related to re-playing the original song and the transcription (play wav, play transcription, play both, resize-play-region, move-play-region).

In total, there were 53,411 user actions on the interface. The average number of user actions per clip is 60.3, with some workers making as few as 1 edit, and other workers making as many as 920 edits for a single clip. Figure 5 shows that the worse the initial automatic transcription, the more actions the Turkers took to make the corrections, which is quite intuitive.

Table 2 shows the average occurrence of each worker action performed over all transcription tasks. By far, the most frequently taken actions were "change pitch" (changing the pitch of a note up or down) and "play both" (playing both the transcription and the audio clip at the same time). The prevalence of the "change pitch, then replay" interaction potentially reflects workers' general difficulty in determining whether two pitches match. It also may reflect the fact that the interface currently allows only a semitone change at a time. Changing the onset and offset of a note occurred less frequently than changing the pitch, but much more frequently than adding or deleting notes. This behavior could indicate that workers are more inclined to modify a note that exists already in lieu of adding a new note. Together, the results suggest that pitch-matching may be the most challenging aspect of the task, and that workers may have an inherent bias to keep the number of notes in the automated transcription constant, while focusing instead on adjusting the pitch, offset, and onset of existing notes.

6. CONCLUSION

In this paper we introduced Ensemble, a semi-automated system that leverages both algorithms and crowd to perform melody transcription. We reported the characteristics, performance and user-behavior pattern of a non-expert crowd of 105 workers for this complex task. For our experiment, workers were able to improve the initial transcription if it was poor, but found it hard to improve a transcription that was already mostly correct. Despite the crowd workers' sentiment that melody transcription is a difficult task, they also feel that it is a fun and interesting task that can hold their attention. We discover that there is indeed a correlation between the music expertise level of a worker and the F-measure performance of their transcription. Many workers commented on the fact that pitch-matching, while being the most frequent action, is also the most challenging aspect of the task.

In the future we plan to breakdown the results by onset, offset, and pitch-only performance, with the goal of gaining further insight into the strengths and weaknesses of the crowdsourced annotations. Furthermore, currently all transcriptions are evaluated against annotations from a single expert. Since the task is somewhat subjective, we plan to collect additional expert annotations and evaluate expert agreement. This would provide a glass ceiling on the performance we can expect from the untrained crowd. We also plan to develop an aggregation algorithm that collates each worker's contribution to create a single (improved) transcription, investigate whether certain granularities (e.g., shorter/longer clips) and decompositions (e.g., having workers specialize in a particular subtask, such as pitch changes) of the task can produce superior transcriptions, and develop new ways to identify the skillful transcribers in the crowd and incentivize them to perform the task.

7. REFERENCES

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

[2] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *ISMIR*, pages 591–596, Miami, USA, 2011.

[3] A. S. Bregman, P. A. Ahad, and J. Kim. Resetting the pitch-analysis system. 2. role of sudden onsets and offsets in the perception of individual components in a cluster of overlapping tones. *J. Acoust. Soc. Am.*, 96(5):2694–2703, 1994.

[4] T. De Clercq and D. Temperley. A corpus analysis of rock harmony. *Popular Music*, 30(1):47–70, 2011.

[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] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE TASLP*, 18(6):1643–1654, 2010.

[7] J. Fritsch. High quality musical audio source separation. Master's thesis, UPMC / IRCAM / Telecom ParisTech, 2012.

[8] B. Fuentes, R. Badeau, and G. Richard. Blind harmonic adaptive decomposition applied to supervised source separation. In *EUSIPCO*, pages 2654–2658, 2012.

[9] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel. Evaluation of a score-informed source separation system. In *ISMIR*, pages 219–224, 2010.

[10] E. Gómez, F. Cañadas, J. Salamon, J. Bonada, P. Vera, and P. Cabañas. Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing. In *ISMIR*, pages 601–606, 2012.

[11] M. Goto. Development of the RWC music database. In *ICA*, pages I–553–556, 2004.

[12] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. Songle: A Web Service For Active Music Listening Improved by User Contributions. In *ISMIR*, pages 311–316, 2011.

[13] S. W. Hainsworth and M. D. Macleod. The automated music transcription problem, 2003.

[14] E. J. Humphrey, J. Salamon, O. Nieto, J. Forsyth, R. Bittner, and J. P. Bello. JAMS: A JSON annotated music specification for reproducible MIR research. In *ISMIR*, pages 591–596, 2014.

[15] H. Kirchhoff, S. Dixon, and A. Klapuri. Shift-variant non-negative matrix deconvolution for music transcription. In *IEEE ICASSP*, pages 125–128, 2012.

[16] A. Laaksonen. Semi-Automatic Melody Extraction Using Note Onset Time and Pitch Information From Users. In *SMC*, pages 689–694, 2013.

[17] E. Law and L. von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *CHI*, pages 1197–1206, 2009.

[18] R. Mason and S. Harrington. Perception and detection of auditory offsets with single simple musical stimuli in a reverberant environment. In *AES*, pages 331–342, 2007.

[19] E. Molina, L. J. Tardón, A. M. Barbancho, and I. Barbancho. SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve. *IEEE TASLP*, 23(2):252–263, 2015.

[20] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE TASLP*, 20(4):1118–1133, 2012.

[21] C. Raffel and D. P. W. Ellis. Intuitive analysis, creation and manipulation of midi data with pretty_midi. In *ISMIR*, 2014.

[22] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *ISMIR*, pages 367–372, 2014.

[23] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Columbia University, 2016.

[24] D. Retelny, S. Robaszkiewicz, A. To, W. Lasecki, J. Patel, N. Rahmati, R. Doshi, M. Valentine Melissa, and M. Bernstein. Expert crowdsourcing with flash teams. In *UIST*, pages 75–85, 2014.

[25] S. Rubin and M. Agrawala. Generating emotionally relevant musical scores for audio stories. In *UIST*, pages 439–448, 2014.

[26] R. Ryan. Control and information in the interpersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 42:450–461, 1982.

[27] M. Ryynänen and A. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

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

[29] P. Smaragdis and G.J. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 69–72, 2009.

[30] M. Staffelbach, P. Sempolinski, D. Hachen, A. Kareem, T. Kijewski-Correa, D. Thain, D. Wei, and G. Madey. Lessons learned from an experiment in crowdsourcing complex citizen engineering tasks with amazon mechanical turk. *CoRR*, abs/1406.7588, 2014.

[31] L. Su and Y.-H. Yang. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *ISMIR*, pages 221–233, 2015.

[32] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. In *ISMIR*, 2007.

[33] J. Urbano, J. Morato, M. Marrero, and D. Mart. Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks. In *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*, pages 9–16, 2010.