INTEGRATION AND QUALITY ASSESSMENT OF HETEROGENEOUS CHORD SEQUENCES USING DATA FUSION

Hendrik Vincent Koops [1] **W. Bas de Haas** [2] **Dimitrios Bountouridis** [1] **Anja Volk** [1]

[1] Department of Information and Computing Sciences, Utrecht University, the Netherlands
{h.v.koops, d.bountouridis, a.volk}@uu.nl
[2] Chordify, the Netherlands {bas}@chordify.net

ABSTRACT

Two heads are better than one, and the many are smarter than the few. Integrating knowledge from multiple sources has shown to increase retrieval and classification accuracy in many domains. The recent explosion of crowd-sourced information, such as on websites hosting chords and tabs for popular songs, calls for sophisticated algorithms for data-driven quality assessment and data integration to create better, and more reliable data. In this paper, we propose to integrate the heterogeneous output of multiple automatic chord extraction algorithms using data fusion. First we show that data fusion creates significantly better chord label sequences from multiple sources, outperforming its source material, majority voting and random source integration. Second, we show that data fusion is capable of assessing the quality of sources with high precision from source agreement, without any ground-truth knowledge. Our study contributes to a growing body of work showing the benefits of integrating knowledge from multiple sources in an advanced way.

1. INTRODUCTION AND RELATED WORK

With the rapid growth and expansion of online sources containing user-generated content, a large amount of conflicting data can be found in many domains. For example, different encyclopediæ can provide conflicting information on the same subject, and different websites can provide conflicting departure times for public transportation. A typical example in the music domain is provided by websites offering data that allows for playing along with popular songs, such as tabs or chords. These websites often provide multiple, conflicting chord label sequences for the same song. The availability of these large amounts of data poses the interesting problem of how to combine the knowledge from different sources to obtain better, and more reliable data. In this research, we address the problem of finding the most appropriate chord label sequence for a piece out of conflicting chord label sequences. Because the correctness of chord labels is hard to define (see e.g. [26]), we define "appropriate" in the context of this research as agreeing with a ground truth. An example of another evaluation context could be user satisfaction.

A pivotal problem for integrating data from different sources is determining which source is more trustworthy. Assessing the trustworthiness of a source from its data is a non-trivial problem. Web sources often supply an external quality assessment of the data they provide, for example through user ratings (e.g. three or five stars), or popularity measurements such as search engine page rankings. Unfortunately, Macrae et al. have shown in [18] that no correlation was found with the quality of tabs and user ratings or search engine page ranks. They propose that a better way to assess source quality is to use features such as the agreement (concurrency) between the data. Naive methods of assessing source agreement are often based on the assumption that the value provided by the majority of the sources is the correct one. For example, [1] integrates multiple symbolic music sequences that originate from different optical music recognition (OMR) algorithms by picking the symbol with the absolute majority at every position in the sequences. It was found that OMR may be improved using naive source agreement measures, but that substantial improvements may need more elaborate methods.

Improving results by combining the power of multiple algorithms is an active research area in the music domain, whether it is integrating the output of similar algorithms [28], or the integration of the output of different algorithms [15], such as the integration of features into a single feature vector to combine the strengths of multiple feature extractors [12, 19, 20]. Nevertheless, none of these deal with the integration and quality assessment of heterogeneous categorical data provided by different sources.

Recent advancements in data science have resulted in sophisticated data integration techniques falling under the umbrella term *data fusion*, in which the notion of source agreement plays a central role. We show that data fusion can achieve a more accurate integration than naive methods by estimating the trustworthiness of a source, compared to the more naive approach of just looking at which value is the most common among sources. To our knowledge no research into data fusion exists in the music domain. Research in other domains has shown that data fusion is capable of assessing correct values with high precision, and significantly outperforms other integration methods [7,25].

In this research, we apply data fusion to the problem of finding the most appropriate chord label sequence for a piece by integrating heterogeneous chord label sequences. We use a method inspired by the ACCUCOPY model that was introduced by Dong et al. in [7, 8] to integrate conflicting databases. Instead of databases, we propose to integrate chord label sequences. With the growing amount of crowd-sourced chord label sequences online, integration and quality assessment of chord label sequences are important for a number of reasons. First, finding the most appropriate chord labels from a large amount of possibly noisy sources by hand is a very cumbersome process. An automated process combining the shared knowledge among sources solves this problem by offering a high quality integration. Second, to be able to rank and offer high quality data to their users, websites offering conflicting chord label data need a good way to separate the wheat from the chaff. Nevertheless, as was argued above, both integration and quality assessment have shown to be hard problems.

To measure the quality of chord label sequence integration, we propose to integrate the outputs of different MIREX Audio Chord Estimation (ACE) algorithms. We chose this data, because it offers us the most reliable ground truth information, and detailed analysis of the algorithms to make a high quality assessment of the integrated output. Our hypothesis is that through data fusion, we can create a chord label sequence that is significantly better in terms of comparison to a ground truth than the individual estimations. Secondly, we hypothesize that the results of integrated chord label sequences have a lower standard deviation on their quality, hence are more reliable.

Contribution. The contribution of this paper is threefold. First, we show the first application of data fusion in the domain of symbolic music. In doing so, we address the question how heterogeneous chord label sequences describing a single piece of music can be combined into an improved chord label sequence. We show that data fusion outperforms majority voting and random picking of source values. Second, we show how data fusion can be used to accurately estimate the relative quality of heterogeneous chord label sequences. Data fusion is better at capturing source quality than the most frequently used source quality assessment methods in multiple sequence analysis. Third, we show that our purely data-driven method is capable of capturing important knowledge shared among sources, without incorporating domain knowledge.

Synopsis. The remainder of this paper is structured as follows: Section 2 provides an introduction to data fusion. Section 3 details how integration of chord label sequences using data fusion is evaluated. Section 4 details the results of integrating submissions of the MIREX 2013 automatic chord extraction task. The paper closes with conclusions and a discussion, which can be found in Section 5.

2. DATA FUSION

We investigate the problem of integrating heterogeneous chord label sequences using data fusion. Traditionally, the goal of data fusion is to find the correct values within autonomous and heterogeneous databases (e.g. [9]). For example, if we obtain meta-data (fields such as year, composer, etc) from different web sources of the song "Black Bird" by The Beatles, there is a high probability that some sources will contradict each other on some values. Some sources will attribute the composer correctly to "Lennon - McCartney", but others will provide just "McCartney", "McCarthey", etc. Typos, malicious editing, data corruption, incorrectly predicted values, and human ignorance are some of the reasons why sources are hardly ever error-free.

Nevertheless, if we assume that most of the values that sources provide are correct, we can argue that values that are shared among a large amount of sources are often more *probable* to be correct than values that are provided by only a single source. Under the same assumption, we can also argue that sources that agree more with other sources are more *accurate*, because they share more values that are likely to be correct. Therefore, if a value is provided by only a single but very accurate source, we can prefer it over values with higher probabilities from less accurate sources, the same way we are more open to accepting a deviating answer from a reputable source in an everyday discussion.

In the above examples, we assume that each source is independent. In real-life this is rarely the case: information can be copied from one website to the other, students repeat what their teacher tells them and one user can enter the same values in a database twice, which can lead to inappropriate values being copied by a large number of sources: *"A lie told often enough becomes the truth"* (Lenin [1]) [8]. Intuitively, we can predict the *dependency* of sources from their sharing of inappropriate values. In general, inappropriate values are assumed to be uniformly distributed, which implies that sharing a couple of identical inappropriate values is a rare event. For example, the rare event of two students sharing a number of identical inappropriate answers on an exam is indicative of copying from each other. Therefore, by analyzing which values with low probabilities are shared between sources, we can calculate a probability of their dependence.

In this research, instead of using databases, we address these issues through data fusion on heterogeneous chord label sequences. Our goal is to take heterogeneous chord label sequences of the same song and create a chord label sequence that is better than the individual ones. We take into account: 1) the *accuracy* of sources, 2) the probabilities of the values provided by sources, and 3) the probability of *dependency* between sources. In the following sections, we refer to different versions of the same song as *sources*, each providing a sequence of values called *chord labels*. See Table 1 for an example, showing four sources ($S_{0...3}$), each providing a sequence of three chord labels, and FUSION, an example of data fusion output.

[1] Ironically, this quote's origin is unclear, but *most* sources cite Lenin.

S$_0$	C:maj	A:min	A:min	F:maj
S$_1$	C:maj	F:maj	G:maj	F:maj
S$_2$	C:maj	F:maj	A:min	D:min
S$_3$	C:maj	F:maj	A:min	D:min
MV	C:maj	F:maj	A:min	?
DF	C:maj	F:maj	A:min	D:min

Table 1: Example of four sources S$_{(0\ldots3)}$ providing different chord label sequences for the same song. DF shows an example output of data fusion on these sources. DF is identical to majority vote (MV) on the first three chord labels. For the last chord label, DF chooses D:min by taking into account source accuracy, while majority vote would randomly pick either F:maj or D:min.

2.1 Source Accuracy

By taking into account the accuracy of a source, we can deal with issues that arise from simple majority voting. For example in Table 1, the final chord labels in the sequence (F:maj and D:min) are provided by the same number of sources. Solving which chord to choose here would require picking randomly one of the two, or using auxiliary knowledge such as harmony theory to make a good choice.

Another problem is that sometimes a source can provide an appropriate chord label that contradicts all other sources. Majority vote would assign the lowest probability to this chord, although it might come from a source that overall agrees a lot with other sources. Intuitively, we have more trust in a source that we believe is more accurate, which is implemented as follows. The chord labels of a source are weighted according to the overall performance of that source: if a source provides a large number of values that agree with other sources, we consider it to be more accurate and more trustworthy, and vice versa.

The accuracy of a source is defined by Dong et al. in [7] as follows. We calculate source accuracy by taking the arithmetic mean of the probabilities of all chord labels the source provides. As an example, suppose we estimate the probabilities of the chords in Table 1 based on their frequency count (c.q. likelihood). That is, C:maj for the first column is 1, A:min for the second column is $1/4$, etc. Then, if we take the average of the chord label probabilities of the first source in our example of Table 1 we can calculate the source accuracy $A(S_0)$ of S_0 as follows:

$$A(S_0) = \frac{1 + 1/4 + 3/4 + 1/2}{4} = 0.625 \tag{1}$$

In the same way, we can calculate the source accuracies for the other three sources which are 0.625, 0.75 and 0.75 for S$_1$,S$_2$ and S$_3$ respectively.

Assuming that the sources are independent, then the probability that a source provides an appropriate chord label is its source accuracy. Conversely, the probability that a source provides an inappropriate chord is the fraction of the inverse of the source accuracy over all possible inappropriate values n: $\frac{(1-A(S))}{n}$. For example, for major and minor chord labels we have 12 roots and 2 modes, which means that for every correct chord label there are $n = (12 * 2) - 1 = 23$ inappropriate chord labels. With more complex chord labels (sevenths, added notes, inversions), n increases combinatorially.

The chord labels of sources with higher accuracies will be more likely to be selected through the use of *vote counts*, which are used as weights for the probabilities of the chord labels they provide. With n and $A(S_i)$ we can derive a vote count $VS(S_i)$ of a source S_i. The vote count of a source is computed as follows:

$$VS(S_i) = \ln \frac{nA(S_i)}{1 - A(S_i)} \tag{2}$$

Applied to our example, this results in vote counts of 2.62 for S$_0$ and S$_1$, and 2.80 for S$_2$ and S$_3$. The higher vote count for S$_2$ and S$_3$ means that its values are more likely to be appropriate than those of S$_0$ and S$_1$.

2.2 Chord Label Probabilities

After having defined the accuracy of a source, we can now determine which chord labels provided by all the sources are most likely the appropriate labels, by taking into account source accuracy. In the computation of chord label probabilities we take into account a) the number of sources that provide those chord labels and b) the accuracy of their sources. With these values we calculate the vote count $VC(\mathcal{L})$ of a chord label $\mathcal{L}$, which is computed as the sum of the vote counts of its providers:

$$VC(\mathcal{L}) = \sum_{\sigma \in S^{\mathcal{L}}} VS(\sigma) \tag{3}$$

where $S^{\mathcal{L}}$ is the set of all sources that provide the chord label $\mathcal{L}$. For example, for the vote count of F:maj in the last column of the example in Table 1, we take the sum of the vote counts of S$_0$ and S$_1$. For the vote count of D:min we take the sum of the vote counts of S$_2$ and S$_3$. To calculate chord label probabilities from chord label vote counts, we take the fraction of the chord label vote count and the chord label vote counts of all possible chord labels (D):

$$P(\mathcal{L}) = \frac{exp(VC(\mathcal{L}))}{\Sigma_{l \in D}\ exp(VC(l))} \tag{4}$$

Applied to our example from Figure 1, we see that solving this equation for F:maj results in a probability of $P(\texttt{F:maj}) \approx 0.39$, and for D:min results in a probability of $P(\texttt{D:min}) \approx 0.56$. Instead of having to choose randomly as would be necessary in a majority vote, we can now see that D:min is more probable to be the correct chord label, because it is provided by sources that are overall more trustworthy.

2.3 Source Dependency

In the sections above we assumed that all sources are independent. This is not always the case when we deal with real-world data. Often, sources derive their data from a common origin, which means there is some kind of dependency between them. For example, a source can copy chord labels from another source before changing some labels, or some Audio Chord Estimation (ACE) algorithm can estimate multiple (almost) equal chord label sequences with different parameter settings. This can create a bias in computing appropriate values. To account for the bias that can arise from source dependencies, we weight the values of sources we suspect to have a dependency lower. In a sense, we award independent contributions from sources and punish values that we suspect are dependent on other sources.

In data fusion, we can detect source dependency directly from the data by looking at the amount of shared uncommon (rare) chord labels between sources. The intuition is that sharing a large number of uncommon chord labels is evidence for source dependency. With this knowledge, we can compute a weight $I(S_i, \mathcal{L})$ for the vote count $VC(\mathcal{L})$ of a chord label $\mathcal{L}$. This weight tells us the probability that a source S_i provides a chord label $\mathcal{L}$ independently.

2.4 Solving Catch-22: Iterative Approach

The chord label probabilities, source accuracy and source dependency are all defined in terms of each other, which poses a problem for calculating these values. As a solution, we initialize the chord label probabilities with equal probabilities and iteratively compute source dependency, chord label probabilities and source accuracy until the chord label probabilities converge or oscillation of values is detected. The resulting chord label sequence is composed of the chord labels with the highest probabilities.

For detailed Bayesian analyses of the techniques mentioned above we refer to [7,10]. With regard to the scalability of data fusion, it has been shown that DF with source dependency runs in polynomial time [7]. Furthermore, [17] propose a scalability method for very large data sets, reducing the time for source dependency calculation by two to three orders of magnitude.

3. EXPERIMENTAL SETUP

To evaluate the improvement of chord label sequences using data fusion we use the output of submissions to the Music Information Retrieval Evaluation eXchange (MIREX) Audio Chord Estimation (ACE) task. For the task, participants extract a sequence of chord labels from an audio music recording. The task requires the estimation chord labels sequences that include the full characterization of chord labels (root, quality, and bass note), as well as their chronological order, specific onset times and durations.

Our evaluation uses estimations from twelve submissions for two Billboard datasets (Section 3.1). Each of these estimations is sampled at a regular time interval to make them suitable for data fusion (Section 3.2). We transform the chord labels of the sampled estimations to different representations (root only, major/minor and major/minor with sevenths) (Section 3.3) to evaluate the integration of different chord types. The sampled estimations are integrated using data fusion per song. To measure the quality of the data fusion integration, we calculate the Weighted Chord Symbol Recall (WCSR) (Section 3.4).

3.1 Billboard datasets

We evaluate data fusion on chord label estimations for two subsets of the Billboard dataset[2], which was introduced by Burgoyne et al. in [3]. The Billboard dataset contains time-aligned transcriptions of chord labels from songs that appeared in the *Billboard* “Hot 100” chart in the United States between 1958 and 1991. All transcriptions are annotated by trained jazz musicians and verified by independent music experts. For the MIREX 2013 ACE task, two subsets of the Billboard dataset were used: the 2012 Billboard set (BB12) and the 2013 Billboard (BB13) set. BB12 contains chord label annotations for 188 songs, corresponding to entries 1000—1300 in the Billboard set. BB13 contains the annotations for 188 different songs: entries 1300—1500.

Twelve teams participated for both datasets, some with multiple submissions: CB3 & CB4 [5], CF2 [4], KO1 & KO2 [16], NG1 & NG2 [13], NMSD1 & NMSD2 [21], PP3 & PP4 [22], and SB [27]. Their submissions are used to evaluate data fusion, for which the Billboard annotations serve as a ground truth.

3.2 Sampling

The MIREX ACE task requires teams to not only estimate *which* chord labels appear in a song, but also *when* they appear. Because of differences in approaches, timestamps of the estimated chord labels do not necessarily agree between teams. This is a problem for data fusion, which expects an equal length and sampling rate of the sources that will be integrated. As a solution, we sample the estimations at a regular interval.

In the past, MIREX used a 10 millisecond sampling approach to calculate the quality of an estimated chord label sequence. Since MIREX 2013, the ground-truth and estimated chord labels are viewed as continuous segmentations of the audio [23]. Because of our data constraint, we use the pre-2013 10 millisecond sampling approach. An initial evaluation using different sampling frequencies in the range 0.1 millisecond to 0.5 seconds, we found only minor differences in data fusion output. The estimated chord label sequences are sampled per song from each team, and used as input to the data fusion algorithm.

3.3 Chord Types

The MIREX ACE task is evaluated on different chord types. To accurately compare our results with those of the teams, and to investigate the effect of integrating different chord types, we follow the chord vocabulary mappings that were introduced by [23] and are standardized in the MIREX evaluation. We map the sampled sequences of estimated chord labels into three chord vocabularies before applying data fusion: root notes only (R), major/minor only chords (MM), and major/minor with sevenths (MM7).

Note that the MIREX 2013 evaluation also includes major/minor with inversions and major/minor seventh chords with inversions. Since there are only two teams that estimated inversions we did not take these into account in our evaluation.

3.4 Evaluation

From the data fusion output sequences for all songs, we calculate the Weighted Chord Symbol Recall (WCSR). The WCSR reflects the proportion of correctly labeled chords in a single song, weighted by the length of the song [14, 23]. To measure the improvement of data fusion, we compare

[2] available from http://ddmal.music.mcgill.ca/billboard

its WCSR with the WCSR of the best scoring team. In addition to data fusion, we compute baseline measurements. We compare the data fusion results with a majority vote (MV) and random picking (RND) technique.

For MV we simply take the most frequent chord label every 10 milliseconds. In case multiple chord labels are most frequent, we randomly pick from the most frequent chord labels. For the example in Table 1, the output would be either `C:maj, F:maj, A:min, F:maj` or `C:maj, F:maj, A:min, D:min`. For RND we select a chord from a random source every 10 milliseconds. For the example in Table 1, RND essentially picks one from 4^4 possible chord label combinations by picking a chord label from a randomly chosen source per column.

4. RESULTS

We are interested in obtaining *improved*, *reliable* chord sequences from *quality assessed* existing estimations. Therefore, we analyze our results in three ways. Firstly, to measure improvement, we show the difference in WCSR between the best scoring team and RND, MV and DF. This way, we can analyze the performance increase (or decrease) for each of these integration methods. The differences are visualized in Figure 1 for the BB12 and BB13 datasets. For each of the three methods, it shows the difference in WCSR for root notes R major/minor only chords MM, and major/minor + sevenths chords (MM7). For detailed individual results an analyses of the teams on both datasets, we refer to [2] and MIREX. [3]

Secondly, to measure the reliability of the integrations, we analyze the standard deviation of the scores of MV and DF. We leave RND out of this analysis because of its poor results. The ideal integration should have 1) a high WCSR and 2) a low standard deviation, because this means that the integration is 1) good and 2) reliable. Table 2 shows the difference with the average standard deviation of the teams. Sections 4.1 - 4.2 report the results in WCSR difference and standard deviation.

Thirdly, in Section 4.3 we analyze the correlation between source accuracy and WCSR, and compare the correlation with other source quality assessments. These correlations will tell us to which extent DF is capable of assessing the quality of sources compared to other, widely used multiple sequence analysis methods.

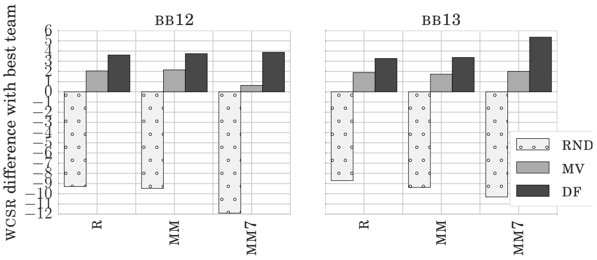

Figure 1: Difference in WCSR with best team for random picking (RND), majority vote (MV) and data fusion (DF). R = root notes, MM = major/minor chords and MM7 = major/minor + sevenths.

	BB12			BB13		
	R	MM	MM7	R	MM	MM7
DF	**-2.5**	**-2.8**	**-2.2**	**-0.5**	**-0.9**	**-1.8**
MV	-1.4	-1.8	-0.97	-0.3	-0.4	-1

Table 2: Difference in standard deviation for DF and MV compared to the average standard deviation of the teams. Lower is better, best values are bold.

4.1 Results of Integrating R, MM and MM7

The left hand sides of the triple-bar groups in Figure 1 show that for both BB12 and BB13, RND performs the worst among RND, MV and DF. RND decreases the WCSR between 8.7% and 12% point, compared to the best performing teams (CB3 and KO1 for BB12 and BB13 respectively) for all chord types. This means that picking random values from sources does not capture shared knowledge in a meaningful way. The middle bars in Figure 1 show that MV integrates knowledge better than RND. MV moderately improves the best algorithm with a difference between 0.6% and 2.1% point.

The right hand sides of the bar groups in Figure 1 show that in both datasets and in all chord types, DF outperforms all other methods with an increase between 3.6% point and 5.4% point compared to the best team. We tested the scores of RND, MV and DF and the best performing teams using a Friedman test for repeated measurements, accompanied by Tukeys Honest Significant Difference tests for each pair of algorithms. We find that DF significantly outperforms the best submission, RND and MV on all datasets on all datasets ($p < 0.01$). These results combined show that DF is capable of capturing knowledge shared among sources needed to outperform all other methods.

In Table 2, we find that for both BB12 and BB13, both MV and DF decrease the standard deviation compared to the average standard deviation of the teams. In fact, we find that DF outperforms MV, improving the standard deviation by a factor two compared to MV. Together, these results mean that on average, DF creates the best sequences with the least errors for all datasets and all chord types.

4.2 Influence of Chord Types on Integration

The results detailed above show that DF is not only capable to significantly outperform all other tested methods on all tested chord labels types, but also produces the most reliable output, because of the low standard deviation.

Comparing the RND, MV and DF results between chord types in Figure 1, we see that the WCSR of RND decreases with a larger chord vocabulary. Because specificity increases the probability of random errors for any algorithm, the probability that RND will pick a good chord label randomly goes down with an increase of the chord vocabulary. For MV, we see that the results are somewhat stable with an increase of the chord vocabulary. Nevertheless, MV is also sensitive for randomly matching chord labels, which explains the drop in accuracy for MM7 for BB13 on the left hand side of Figure 1. Most interestingly, we observe that the performance of DF increases with a larger chord vocabulary. The explanation is that specificity helps DF to separate good sources from bad sources. With a larger chord vocabulary, sources will agree with each other on

[3] http://www.music-ir.org/mirex/wiki/2013:MIREX2013_Results

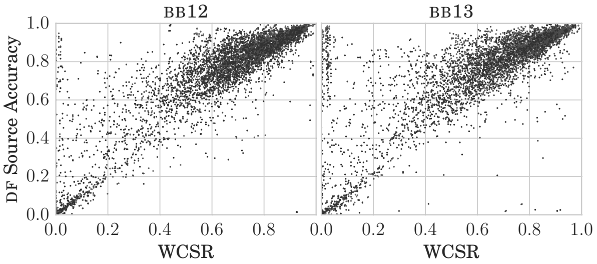

Figure 2: Correlation between WCSR and source accuracy. Plotted are R, MM and MM7. One dot is one estimated chord label sequence for one song from one team.

more specific chord labels, which decreases the probability of unwanted random source agreement.

4.3 Source Quality Assessment

The previous sections show that data fusion is capable of selecting good chord labels from the coherence between the sources, without ground truth knowledge. A pivotal part of data fusion is the computation of source accuracy, which provides a relative score for each source compared to the other sources. There are circumstances in which we are more interested in the estimation of source accuracy than the actual integration of source data. For example, ranking a number of different crowd sourced chord label sequences of the same song obtained from web sources, (e.g. investigated by [18]). Investigating the relationship between source accuracy and the WCSR provides insight whether data fusion is capable of assessing the accuracy of the sources in a way that reflects WCSR. WCSR reflects the quality of the chord sequences and therefore the quality of the algorithm. This relationship is shown in Figure 2, in which the WCSR is plotted against the DF source accuracy.

Initial observation of Figure 2 shows that for both BB12 and BB13, WCSR and source accuracy are distributed along a more or less diagonal line, meaning that a higher WCSR is associated with a higher DF source accuracy, and vice versa. This indicates a strong correlation, which is confirmed by the Spearman's rank correlation coefficient (SRCC). To analyze the relative performance of source quality assessment of DF, we compare its correlation with widely used sequence scoring methods. These are often used in bioinformatics, where sequence ranking is at the root of a multitude of problems. Table 3 compares the SRCC of different similarity scoring methods for BB12 and BB13. The table shows the correlations between WCSR and DF, bigrams (BIGRAM), profile hidden Markov models (PHMM), percentage identity (PID), and neigbor-joining trees (NJT). BIGRAM compares the relative balance of specific character pairs appearing in succession, also known as bigrams. Sequences belonging to the same group should be stochastic products of the same probabilistic model [6]. PHMM turns the sources into a position-specific scoring system by creating a profile with position-probabilities. A source is scored through comparison with the profile of all other sources [11]. PID is the fraction of equal characters divided by the length of the source. NJT is a bottom-up clustering method for the creation of phylogenetic trees, in which the distance from the root is the score [24].

	BB12			BB13		
	R	MM	MM7	R	MM	MM7
DF	**0.87**	**0.85**	**0.82**	**0.77**	**0.77**	**0.76**
BIGRAM	0.18	0.18	0.16	0.2	0.22	0.29
PHMM [4]	0.22	—	—	0.22	—	—
PID	0.18	0.2	0.19	0.25	0.27	0.29
NJT	0.2	0.22	0.21	0.24	0.25	0.27

Table 3: Spearman's rank correlation coefficient (ρ) of WCSR and other source scoring methods. Best performing algorithms are bold. All values are significant with $p < 0.01$.

The table shows that DF source accuracy has the highest correlation with WCSR among all other methods. These results show that data fusion is capable of assessing the quality of the sources without any ground-truth knowledge in a way that is closely related to the actual source quality.

5. DISCUSSION AND CONCLUSION

Through this study, we have shown for the first time that using data fusion, we can integrate the knowledge contained in heterogeneous ACE output to create improved, and more reliable chord label sequences. Data fusion integration outperforms all individual ACE algorithms, as well as majority voting and random picking of source values. Furthermore, we have shown that with data fusion, one can not only generate high quality integrations, but also accurately estimate the quality of sources from their coherence, without any ground truth knowledge. Source accuracy outperforms other popular sequence ranking methods.

Our findings demonstrate that knowledge from multiple sources can be integrated effectively, efficiently and in an intuitive way. Because the proposed method is agnostic to the domain of the data, it could be applied to melodies or other musical sequences as well. We believe that further analysis of data fusion in crowd-sourced data has the potential to provide non-trivial insights into musical variation, ambiguity and perception. We believe that data fusion has many important applications in music information retrieval research and in the music industry for problems relating to managing large amounts of crowd-sourced data.

Acknowledgements We thank anonymous reviewers for providing valuable comments on an earlier draft on this text. H.V. Koops and A. Volk are supported by the Netherlands Organization for Scientific Research, through the NWO-VIDI-grant 276-35-001 to A. Volk. D. Bountouridis is supported by the FES project COMMIT/.

6. REFERENCES

[1] E.P. Bugge, K.L. Juncher, B.S. Mathiesen, and J.G. Simonsen. Using sequence alignment and voting to improve optical music recognition from multiple recognizers. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 405–410, 2011.

[2] J.A. Burgoyne, W.B. de Haas, and J. Pauwels. On comparative statistics for labelling tasks: What can we learn from MIREX ACE 2013. In *Proc. of the 15th*

[4] The MM and MM7 chord label alphabets are too large for the used PHMM application, which only accepts a smaller bioinformatics alphabet.

Conference of the International Society for Music Information Retrieval, pages 525–530, 2014.

[3] J.A. Burgoyne, J. Wild, and I. Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proc. of the International Society for Music Information Retrieval Conference*, volume 11, pages 633–638, 2011.

[4] C. Cannam, M. Mauch, M.E.P. Davies, S. Dixon, C. Landone, K. Noland, M. Levy, M. Zanoni, D. Stowell, and L.A. Figueira. MIREX 2013 entry: Vamp plugins from the centre for digital music, 2013.

[5] T. Cho and J.P. Bello. MIREX 2013: Large vocabulary chord recognition system using multi-band features and a multi-stream hmm. *Music Information Retrieval Evaluation eXchange (MIREX)*, 2013.

[6] M.J. Collins. A new statistical parser based on bigram lexical dependencies. In *Proc. of the 34th annual meeting on Association for Computational Linguistics*, pages 184–191. Association for Computational Linguistics, 1996.

[7] X.L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proc. of the VLDB Endowment*, 2(1):550–561, 2009.

[8] X.L. Dong and F. Naumann. Data fusion: resolving data conflicts for integration. *Proc. of the VLDB Endowment*, 2(2):1654–1655, 2009.

[9] X.L. Dong and D. Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.

[10] X.L. Dong and D. Srivastava. Big data integration. *Synthesis Lectures on Data Management*, 7(1):1–198, 2015.

[11] S.R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[12] R. Foucard, S. Essid, M. Lagrange, G. Richard, et al. Multi-scale temporal fusion by boosting for music classification. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 663–668, 2011.

[13] N. Glazyrin. Audio chord estimation using chroma reduced spectrogram and self-similarity. *Music Information Retrieval Evaluation Exchange (MIREX)*, 2012.

[14] C. Harte. *Towards automatic extraction of harmony information from music signals*. PhD thesis, Department of Electronic Engineering, Queen Mary, University of London, 2010.

[15] A. Holzapfel, M.E.P. Davies, J.R. Zapata, J.L. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9):2539–2548, 2012.

[16] M. Khadkevich and M. Omologo. Time-frequency reassigned features for automatic chord recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 181–184. IEEE, 2011.

[17] X. Li, X.L. Dong, K.B. Lyons, W. Meng, and D. Srivastava. Scaling up copy detection. *arXiv preprint arXiv:1503.00309*, 2015.

[18] R. Macrae and S. Dixon. Guitar tab mining, analysis and ranking. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 453–458, 2011.

[19] A. Meng, P. Ahrendt, and J. Larsen. Improving music genre classification by short time feature integration. In *Acoustics, Speech, and Signal Processing, 2005. Proc..(ICASSP'05). IEEE International Conference on*, volume 5, pages v–497. IEEE, 2005.

[20] A. Meng, J. Larsen, and L.K. Hansen. *Temporal feature integration for music organisation*. PhD thesis, Technical University of Denmark, Department of Informatics and Mathematical Modeling, 2006.

[21] Y. Ni, M. Mcvicar, R. Santos-Rodriguez, and T. De Bie. Harmony progression analyzer for MIREX 2013. *Music Information Retrieval Evaluation eXchange (MIREX).*

[22] J. Pauwels, J-P. Martens, and G. Peeters. The ircamkeychord submission for MIREX 2012.

[23] J. Pauwels and G. Peeters. Evaluating automatically estimated chord sequences. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 749–753. IEEE, 2013.

[24] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[25] N.T. Siebel and S. Maybank. Fusion of multiple tracking algorithms for robust people tracking. In *Computer VisionECCV 2002*, pages 373–387. Springer, 2002.

[26] 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 *Proc. of the International Society for Music Information Retrieval Conference*, volume 11, pages 555–560, 2011.

[27] Nikolaas Steenbergen and John Ashley Burgoyne. Joint optimization of an hidden markov model-neural network hybrid for chord estimation. *MIREX-Music Information Retrieval Evaluation eXchange. Curitiba, Brasil*, pages 189–190, 2013.

[28] C. Sutton, E. Vincent, M. Plumbley, and J. Bello. Transcription of vocal melodies using voice characteristics and algorithm fusion. In *2006 Music Information Retrieval Evaluation eXchange (MIREX)*, 2006.