MAP ADAPTATION TO IMPROVE OPTICAL MUSIC RECOGNITION OF EARLY MUSIC DOCUMENTS USING HIDDEN MARKOV MODELS

Laurent Pugin **John Ashley Burgoyne** **Ichiro Fujinaga**
Centre for Interdisciplinary Research in Music Media and Technology
Schulich School of Music of McGill University
Montréal, Québec, Canada H3A 1E3
{laurent,ashley,ich}@music.mcgill.ca

ABSTRACT

Despite steady improvement in optical music recognition (OMR), early documents remain challenging because of the high variability in their contents. In this paper, we present an original approach using maximum a posteriori (MAP) adaptation to improve an OMR tool for early typographic prints dynamically based on hidden Markov models. Taking advantage of the fact that during the normal usage of any OMR tool, errors will be corrected, and thus ground-truth produced, the system can be adapted in real-time. We experimented with five 16th-century music prints using 250 pages of music and two procedures in applying MAP adaptation. With only a handful of pages, both recall and precision rates improved even when the baseline was above 95 percent.

1 INTRODUCTION

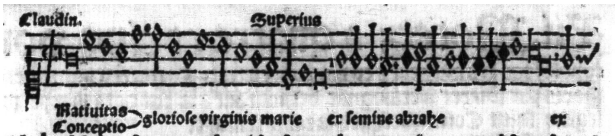
Optical music recognition (OMR) systems create encodings of the musical content in digital images automatically. For libraries, they constitute very promising solutions for building searchable digital libraries of previously inaccessible material, especially historical documents. Using OMR tools in large-scale digitisation project remains a challenge, however [4], mainly because of the inconsistent performance of most OMR tools. Learning-based approaches, such as the one adopted by Gamera [9], are well suited to such projects, but other approaches, e.g., coupling multiple recognisers, have also been considered recently [3].

When performing OMR on early music sources, one major problem is the extremely high variability exhibited in the data. In printed documents, the font shape may vary considerably from one print to another, and the printing techniques of the time as well as the texture of the paper used resulted in frequent printing irregularities. The physical documents are often degraded, introducing various kinds of noise, and the scanning settings (e.g., brightness or contrast) are not necessarily consistent across all documents. Five examples of 16th-century music prints we used for this study, shown in Figure 1, demonstrate the variability in font shape, document degradation, and scanning parameters. Given such a range of documents, there is no guarantee that an OMR system can be trained to perform well on new document, even with a learning-based approach. Furthermore, as a considerable amount of labelled data is required to train a sufficiently reliable system [10], building a system from scratch for every new document encountered is not practical.

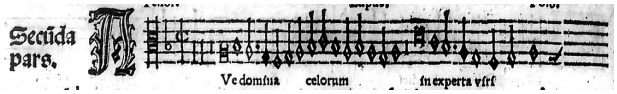
Similar problems have been encountered previously in speech recognition, where the amount of data and time needed to build a recogniser is considerable. One common approach to solve the problem is to use so-called adaptation techniques. With these techniques, when a new speaker has to be recognised, a system that was previously trained on a large set of other speakers can be optimised for the new speaker using only a small set of new examples. One widely used approach for adaptation in speech is maximum a posteriori (MAP) adaptation [8], a technique that has also been applied in other domains such as handwriting recognition (on-line [2] or off-line [14]), audio transcription [7], and video annotation [1].

In OMR, it is difficult to imagine an application where the recognition errors would not have to be hand-corrected before using the output. For example, if the tool is used to build a digital library or to perform music analysis, it is absolutely necessary to have a correct representation of the musical text. This property makes adaptation techniques of prime interest for OMR because the normal usage of any OMR application software provides the hand-corrected data for adaptation and performance improvement. As soon as a page has been recognised and corrected by the user during an OMR process, adaptation can be run so that the subsequent pages will require fewer corrections.

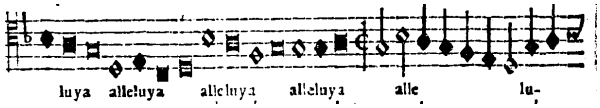
In this paper, we present our experiments in using MAP adaptation in Aruspix, an OMR system for early typographic prints based on hidden Markov models (HMMs) [11]. The main goals of the study were to see whether MAP adaptation works in this context, how the adaptation process has to be organised, and how much data is needed to reap benefits from adaptation. We also compared the results obtained with those obtained when training the system from scratch, i.e., without using an adaptation technique.



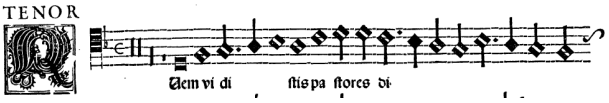
(a) RISM 1528-2 (Attaignant, Paris, 1528)



(b) RISM 1532-10 (Moderne, Lyon, 1532)



(c) RISM V-1421 (Figliuoli di Gardano, Venezia, 1572)



(d) RISM V-1433 (Basa, Roma, 1585)

(e) RISM M-0582 (Le Roy & Ballard, Paris, 1598)

Figure 1: Examples of prints used for the experiments

2 OMR INFRASTRUCTURE

Aruspix provides an infrastructure that handles the complete OMR process. It performs the indispensable preprocessing operations, such as deskewing the image, normalising the size, removing image borders, binarising and cleaning the image, detecting staff positions, and pre-classifying some of the elements (music, ornate letters, lyrics, and title elements). The core of the recognition is performed using HMMs, an original approach to OMR. The Aruspix infrastructure also includes an editor designed especially for early music, which makes it an end-user application as well as a research tool.

The machine learning component of Aruspix (training and recognition) is based on the Torch machine learning library [5]. Recognition is performed using continuous-density HMMs with a 2-pixel sliding window (for a normalised staff height of 100 pixels). At each position, a feature vector of 7 values is extracted [12], and the sequences of feature vectors are then used to build a set of HMMs using embedded training over the whole staff.

3 MAP ADAPTATION

The general schema to enable MAP adaptation in Aruspix is to build, in a preliminary phase, a book-independent (BI) system using as large a set of learning data as possible, taken from many different books. The BI system gives acceptable results in general but is not optimised for any book in particular. When a page has been recognised and corrected by the user, the BI system is then adapted with MAP adaptation using the corrected page as an example. This means that, through usage, a book-dependent (BD) system for the book currently being processed can be obtained very quickly.

When training the HMMs for the BI system, the expectation-maximisation (EM) algorithm is used to determine the parameter vector λ that maximises $P(X|\lambda)$, where X is the observed data, i.e.:

$$\lambda = \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda) \tag{1}$$

The principle of MAP adaptation [6] is to find the parameters λ_M that maximises the posterior probability $P(\lambda|X)$ using the prior knowledge about the model parameter distribution of the already trained model $P(\lambda)$:

$$\lambda_M = \underset{\lambda}{\operatorname{argmax}}\, P(\lambda|X) = \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda)P(\lambda) \tag{2}$$

To obtain a λ_M estimate for the BD model, the EM algorithm is applied, and the value obtained enables the means μ of the BI HMMs to be adapted, while the variances, the transitions, and weights are usually unchanged [14]. EM runs with a heuristic weighting factor τ on the relative importance of the new adaptation data. High values of τ privilege the BI model, while low values privilege the new adaptation data. This weighting factor has to be determined empirically.

4 EXPERIMENTS

For our experiments, we used microfilms of sixteenth-century music prints held at the Marvin Duchow Music Library at McGill University and the Isham Memorial Library at Harvard University. They were scanned as 8-bit greyscale TIFFs at a resolution of 400 dots per inch. The BI system was trained using 457 pages taken from music books produced by printers from Italy, France, Belgium, and Germany between 1529 and 1595. This set of pages was transcribed and represents a total of 2,710 staves and 95,845 characters. Using this model, Aruspix was trained to recognise 220 different musical symbols (note values from *longa* to *semi-fusa*, rests, clefs, accidentals, *custodes*, dots, bar lines, coloured notes, ligatures, etc.).

To experiment with MAP adaptation, we used 5 books printed in France and Italy between 1528 and 1598 (RISM 1528-2, 1532-10, V-1421, V-1433 and M-0582) [13]. For each of them, 50 pages were transcribed and corrected in Aruspix (250 pages in total). We took the first 40 pages for the training set and reserved the 10 remaining pages for the test set. We decided to select the first pages for adaptation because it is in this order that the data would become available in a digitisation workflow, but to verify our results, we performed a 5-fold cross-validation on one of the books, M-0582. In two books, the pages transcribed contain new symbols not represented in the BI model, 10

in M-0582 and 7 in V-1433, which required special treatment.

We experimented with MAP adaptation using cumulative procedures, common when experimenting with this technique in an off-line architecture [14]. Unlike incremental MAP adaptation, which generates new BD models for every page of adaptation data using only the new page and the BD model from the previous page, in cumulative MAP adaptation, the BD model is generated using the BI model and the complete set of adaptation data up to that point. We tried two approaches in particular. The first is using embedded adaptation with the Viterbi algorithm on whole staves, which is similar to embedded training when training HMMs from scratch [11]. We call it the embedded cumulative MAP (EC-MAP) adaptation. The second approach is to adapt the models for each symbol individually, taking the advantage of the fact that our ground-truth data are aligned. This approach, which we call isolated cumulative MAP (IC-MAP) adaptation, is uncommon because in other domains, the adaptation data are not usually aligned. Finally, we trained a new model from scratch for each of the five books, using the data in a cumulative way, in order to compare with the MAP adaptation results.

The MAP factor τ was empirically optimised in both EC-MAP and IC-MAP. The best results were obtained when the factor was decreased as the amount of data increased, reflecting the intuitive assumption that the more data we have for MAP adaptation, the less we need to rely on the original model. During the MAP adaptation process, the new symbols in M-0582 and V-1433 were necessarily trained separately before being inserted into the system.

5 RESULTS

The results were evaluated by calculating recall and precision on the best-aligned subsequence of recognised symbols [10]. We computed a baseline by testing the BI model on the five test sets.

5.1 MAP adaptation vs training from scratch

For all five sets, MAP adaptation improved both recall and precision rates (see tables 1 and 2), even where the baseline was above 95% (V-1421). Using all 40 pages of the training sets, MAP adaptation gives better results than training the models from scratch (TS) for all sets but one. In several cases, training from scratch failed to achieve even the baseline recall or precision. The only book where it yields better results is M-0582, the most degraded book in our set (see figure 1e). The severe degradation may explain why in the end, training from scratch can outperform MAP adaptation.

Table 2 shows the adaptation and training curves for the cross-validated results on M-0582. Other than the fact that training from scratch outperforms MAP adaptation after about 30 pages, these two plots are also representative of the curve shapes we obtained for the other sets. We can see that MAP adaptation improves the results after only a handful of pages (between 5 and 10) and that no further significant improvement is obtained after 20 pages.

Table 1: Recall results with 40 pages

Book	Base.	TS	IC-MAP	EC-MAP
1528-2	84.93	89.67	88.36	**91.61**
1532-10	76.53	87.86	86.24	**89.23**
V-1421	95.82	93.84	96.33	**97.01**
V-1433	86.32	91.70	90.82	**92.62**
M-0582	72.44	**90.26**	86.31	88.44

Table 2: Precision results with 40 pages

Book	Base.	TS	IC-MAP	EC-MAP
1528-2	96.57	95.57	95.77	**97.11**
1532-10	94.24	93.29	95.30	**95.98**
V-1421	97.18	94.95	**97.28**	97.05
V-1433	95.48	95.56	71.18	**97.25**
M-0582	84.54	**93.19**	92.19	90.14

5.2 EC-MAP vs IC-MAP adaptation

Overall, IC-MAP adaptation does not give as good results as EC-MAP adaptation. Nevertheless, it merits consideration because it runs so much faster than the EC-MAP adaptation. Table 3 show the mean adaptation time for both adaptation procedures. On our 2.7 GHz PowerPC G5 processor, IC-MAP takes less than one second per page on average, and it increases linearly as the number of pages increases. In comparison, EC-MAP takes about one minute per page and increases exponentially. For this reason, IC-MAP is the most suitable approach to perform real-time adaptation, e.g., within a real-world digitisation workflow. As soon a page is corrected, the model can be adapted before recognising the next page.

The main drawback to IC-MAP is that it requires a good alignment of the adaptation data. One book in our set (V-1433) had poorly aligned data because it had been printed using a much wider font than the others (see figure 1d). We can see in table 2 that the precision decreased with IC-MAP for the particular book.

6 CONCLUSIONS AND FUTURE WORK

To deal with the high variability in early music documents, we recommend the use of MAP adaptation. For the books

Table 3: Mean adaptation time

# of pages	IC-MAP	EC-MAP
1	$>$ 1 sec	1 min
5	5 sec	7 min
10	9 sec	20 min
20	18 sec	50 min
40	38 sec	1 h 45 min

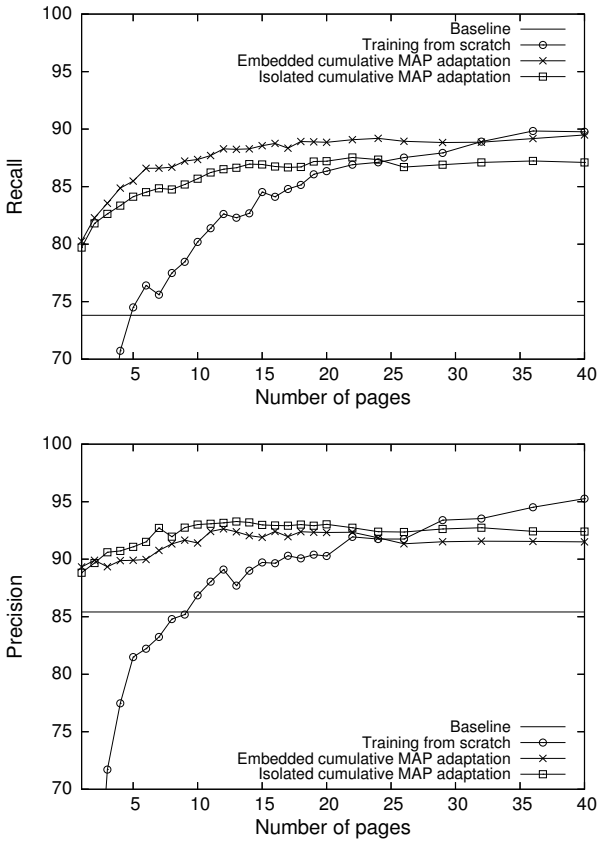


Figure 2: Cross-validated results for M-0582

used for our experiments, MAP adaptation reduced the recognition failures by a factor of nearly two on average without losing precision. Precision was, in fact, increased. Our experiments showed that only a couple of pages are needed to benefit from MAP adaptation. In comparison, when training new models from scratch, at least 30 pages are needed to outperform the results obtained with adaptation. We also presented an original approach in applying MAP adaptation to isolated symbols (IC-MAP), which computes very quickly (about 1 second per page) and could be used in real-time within a typical OMR process. Such an infrastructure will speed up the OMR workflow by exploting the required human editing process to improve machine recognition, although to obtain the greatest benefit, this infrastructure must include efficient user interfaces for error correction. This approach also opens new perspectives for other tasks where the data present high variability, such as music manuscripts or other types of early documents.

7 ACKNOWLEDGEMENTS

We would like to thank the Canada Foundation for Innovation and the Social Sciences and Humanities Research Council of Canada for their financial support. We also would like to thank Marnie Reckenberg for her contribution to the project.

8 REFERENCES

[1] M. Barnard and J.-M. Odobez. Robust playfield segmentation using MAP adaptation. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 610–13, Cambridge, United Kingdom, 2004.

[2] A. Brakensiek, A. Kosmala, and G. Rigoll. Writer adaptation for online handwriting recognition. In *Pattern Recognition: 23rd DAGM Symposium, Munich, Germany, September, 2001, Proceedings*, volume 2191 of *LNCS*, pages 32–37. Springer, Berlin, 2001.

[3] D. Byrd and M. Schindele. Prospects for improving OMR with multiple recognizers. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 41–46, Victoria, Canada, 2006.

[4] G. S. Choudhury, T. DiLauro, M. Droettboom, I. Fujinaga, B. Harrington, and K. MacMillan. Optical music recognition system within a large-scale digitization project. In *Proceedings of the 1st International Conference on Music Information Retrieval*, 2000.

[5] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, IDIAP, Martigny, Switzerland, October 2002.

[6] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–98, 1994.

[7] M. Goto. A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3365–68, Salt Lake City, UT, 2001.

[8] C.-H. Lee, C.-H. Lin, and B.-H. Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Transactions on Signal Processing*, 39(4):806–14, 1991.

[9] K. MacMillan, M. Droettboom, and I. Fujinaga. Gamera: Optical music recognition in a new shell. In *Proceedings of the International Computer Music Conference*, pages 482–85, 2002.

[10] L. Pugin. *Lecture et traitement informatique de typographies musicales anciennes. Un logiciel de reconnaissance de partitions par modèles de Markov cachés*. Ph.D. Dissertation, University of Geneva, 2006.

[11] L. Pugin. Optical music recognition of early typographic prints using hidden Markov models. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 53–56, Victoria, Canada, 2006.

[12] L. Pugin, J. A. Burgoyne, and I. Fujinaga. Goal-directed evaluation for the improvement of optical music recognition on early music prints. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 303–4, Vancouver, Canada, 2007.

[13] Répertoire international des sources musicales (RISM). *Single Prints Before 1800*. Series A/I. Bärenreiter, Kassel, 1971–81.

[14] A. Vinciarelli and S. Bengio. Writer adaptation techniques in HMM based off-line cursive script recognition. *Pattern Recognition Letters*, 23:905–16, 2002.

Many recognition problems can be corrected only using this axis of control, while switch se ings are o en appropriate for entire regions of the page. e other axis of control allows the user to label individual image pixels with the symbol (e.g. sharp, treble clef, quarter rest, slur, augmentation dot, etc.) or primitive (e.g. open note head, ledger line, stem, double beam, stem, etc.) that covers the pixel. e user can re-recognize as many times as is needed, simultaneously adding pixel.Â 2007. MAP Adaptation to Improve Optical Music Recognition of Early Music Documents Using Hidden Markov Models. In ISMIR. 513â€“516. [16] A. Rebelo, G. Capela, and J. S. Cardoso. 2009. Optical Recognition of Music Symbols. International Journal on Document Analysis and Recognition 13 (2009), 19â€“31.