

Pattern Analysis and Applications manuscript No. (will be inserted by the editor)

Hybrid Hidden Markov Models and Artificial Neural Networks for Handwritten Music Recognition in Mensural Notation

Jorge Calvo-Zaragoza, Alejandro H. Toselli, Enrique Vidal

Received: date / Accepted: date

Abstract In this paper we present a hybrid approach using Hidden Markov Models (HMM) and Artificial Neural Networks to deal with the task of Handwritten Music Recognition in Mensural notation. Previous works have shown that the task can be addressed with Gaussian density HMMs that can be trained and used in an end-to-end manner; that is, without prior segmentation of the symbols. However, the results achieved using that approach are not sufficiently accurate to be useful in practice. In this work we hybridize HMMs with deep Multi-Layer Perceptrons (MLP), which lead to remarkable improvements in optical symbol modeling. Moreover, this hybrid architecture maintains important advantages of HMMs such as the ability of properly modeling variable-length symbol sequences through segmentation-free training, and the simplicity and robustness of combining optical models with N -gram language models, which provide statistical *a priori* information about regularities in musical symbol concatenation observed in the training data. The results obtained with the proposed hybrid MLP-HMM approach outperform previous works by a wide margin, achieving symbol-level error rates around 26%, as compared with about 40% reported in previous works.

Keywords Handwritten Music Recognition · Mensural notation · Hidden Markov Models · Artificial Neural Networks · N -gram Language Models

1 Introduction

The preservation of the musical heritage over the centuries makes it possible to go deeper into the study of certain artistic and cultural paradigms. Most of this heritage is stored in cathedrals, archives and music libraries [11]. In addition to issues related to the ownership of the sources, this storage allows the physical preservation of the sources over time; in turn, it also severely limits the access for study and analysis. To improve this situation, significant efforts are being made to digitize relevant music document collections [15]. The resulting digital copies can then be easily accessed and studied without compromising the integrity of the original sources.

Nevertheless, digitization is not enough to exploit the important potential of this heritage. To make the most out of it, the *music content* itself must be transcribed into a structured format that can be easily processed by a computer. This would open possibilities of great interest for the musicological community such as content-based search and digital editing, as well as large-scale musicological analysis by means of computational tools. Of course, this transcription process can be done manually, but the cost of this option is highly prohibitive given the huge size of the collections of interest. In this context, systems for automatic transcription of music manuscripts would clearly constitute very valuable tools [1].

This kind of systems — usually referred as to Optical Music Recognition (OMR), or Handwritten Music Recognition (HMR) when working on handwritten scores — import the image of a musical score and are expected to export its content to symbolic formats such as MusicXML or MEI, to name a few.

OMR/HMR systems, regardless of the approach used to achieve their objective, can be very varied due to the differences over time amongst musical notations, document layouts, or printing mechanisms. In this work, we focus on the development of HMR systems for Mensural music, one of the most commonly used notations during Renaissance. At present, there are millions of manuscripts of this type that remain to be transcribed. Obviously, each manuscript collection may have its own particularities (such as the handwriting style), but the approach developed in this work provides a common formulation to all of them.

Traditionally, HMR systems have followed a two-stage work-flow: segmentation, in which the musical symbols are isolated in the image; and recognition, in which a meaningful label is assigned to each symbol. In contrast, here we adopt a holistic approach, where both stages are integrated and performed simultaneously. To this end, we follow the work reported in [7, 8] and resort to the use of Hidden Markov Models (HMM), which are very convenient to model sequential data, such as the sequences of music primitives (notes) in music notation. However, the optimization strategies of these models pursue a generative objective and, therefore, their potential for discriminative recognition is limited to some extent.

In this work we extend the use of HMMs for HMR by studying a hybrid approach where Artificial Neural Networks are trained to discriminate hidden HMM states in the decoding stage. As will be reported, this approach leads to remarkable optical symbol modeling improvements, while maintaining important advantages of HMMs. In particular, it enables proper modeling of variable-length symbol sequences, segmentation-free training, and simple and robust combination of optical models with N -gram language models. These language models, which are also trained from ground-truth data, provide statistical *a priori* information about expected regularities in musical symbol concatenation, which further help improving recognition accuracy.

These contributions lead us to present the first HMR system for Mensural notation that is trained in an end-to-end manner — without any type of information about the location of the elements in the input staves — and achieves results which are sufficiently precise to be useful in practice. More specifically, we improve the performance of previous works from about 40% symbol error rate to 26%, under identical experimental conditions.

The rest of the article is structured as follows: Section 2 presents the current state of the art through a review of previous works; Section 3 provides information needed in the remaining sections, such as an introduction to Mensural notation, the corpus considered and the image pre-processing adopted; Section 4 describes the HMM framework for HMR and its subsequent hybridization with Neural Networks, as well as the use of statistic language models; Section 5 reports the experimental results obtained, along with analysis and discussion; finally, Section 6 concludes this work, pointing out promising lines of future research.

2 Background

Optical recognition of music notation is a challenge for which it is often claimed that there exist no successful approaches. This is a rather general statement because different levels of difficulty are found depending on factors such as the notation type (modern Western, Mensural, Neumatic, etc.) or the engraving mechanism (handwritten or printed).

This task has been commonly addressed through a series of independent stages that work on different parts of the problem [25]. From a morphological point of view, music notation hardly has what we might consider low-level entities, like phonemes in speech or characters in text, but rather isolated music symbols. This may explain why most previous approaches consider by default that symbol segmentation should be an initial step. Correspondingly, the majority of the research carried out so far has focused on a staff-line removal stage [12] (because staff lines are usually considered an important obstacle for music symbol segmentation), followed by symbol segmentation, and classification of isolated musical symbols [18]. However, symbol segmentation is often difficult, especially in the case of images of handwritten music. It becomes particularly difficult to distinguish between relevant small elements from noise and other artifacts caused by document preservation problems and possible lack of image quality. In addition, some handwritten music symbols are divided into smaller primitives that in the segmentation-based approach are detected separately, and so it is necessary to post-process the results of symbol recognition to assemble the actual music notation. Clearly, the heuristics underlying these solutions for a particular type of documents hardly generalize to other, even similar, mu-

sical documents, and often result in diverse kinds of failures [25].

Not only symbol segmentation is troublesome; it is even more problematic to achieve robust and accurate recognition results from the segmented image patches. In such a case, considering the symbols isolatedly makes it difficult or impossible to benefit from any contextual (“linguistic”) information provided by the surrounding symbols. Few works have dealt so far with the whole recognition task, except for solutions to complete recognition of simple types of music notation, namely square notation [24] or printed Mensural notation [6].

In this work, we are interested in addressing the HMR problem for Mensural notation, mainly used in early handwritten scores. As introduced by the work of Pugin [22], and extended in [7, 8] for handwritten scores, an approach based on Hidden Markov Models (HMM) is certainly interesting in this context. HMM have traditionally been used in tasks with a similar formulation to HMR like text or speech recognition [20], and they continue to be a reference in many disparate duties and scenarios [27, 30].

In our case, the use of HMM allows context-aware recognition of full staff images, as well as a holistic training without depending on any kind of previous symbol segmentation. It also avoids having to rely on aforementioned complex multi-stage pipelines, with many hand-crafted heuristics, which are likely to fail to generalize adequately. Furthermore, producing ground-truth for such holistic approaches is much less demanding, thereby significantly reducing the cost of dealing with new manuscript collections.

The work described in [7, 8] show the feasibility of these ideas for HMR. However, results reported in these papers were still not sufficiently good to be useful in practice. Here we go further in this direction by significantly improving optical modeling and training. To this end, we introduce a hybrid approach based on HMM and Artificial Neural Networks (ANN). HMM naturally model the sequentiality of music notation and the holistic nature of the overall problem, while the ANN provides discriminative modeling and training which is important to effectively deal with the symbol recognition problem.

The weakness of HMM as regards discriminative tasks has been widely discussed in the literature, and that is why it can be found works that hybridize HMM with other schemes [2, 3]. In our case we consider the use of Multi-Layer Perceptrons (MLP), whose probabilistic interpretation fits perfectly well when combining it with HMM. A simi-

lar idea was successfully used for Handwritten Text Recognition [10], but its (promising) use for HMR remains unexplored so far.

3 Preliminaries

Basic concepts and elements of Mensural notation are outlined in this section, followed by a description of the dataset used in this work.

3.1 Mensural Notation

The work presented here deals with manuscripts written in Mensural notation, specifically with sources under the Pan-Hispanic framework of the 17th century. Although this type of Mensural notation is generally considered as an appendix of the European Mensural notation, the Pan-Hispanic context of that time presented a particular development of the musical activity that was almost totally under the control of the ecclesiastical state. This fostered, among other things, the massive use of handwritten copies instead of printed ones. These copies, that are historically considered of greater relevance, arise the necessity of developing successful HMR systems.

This context also caused the development of particular graphic codes that favored the live readability of music. A set of representative symbols from this notation is depicted in Table 1. Among the particularities mentioned above, it should be noted that the color of the note-heads does not change the duration of the sound, as it does in modern notations, but is used to indicate certain particularities of the rhythm. As in almost any music notation, the meaning of most musical symbols relies on two geometrical informations: *shape* and *height* (vertical position of the symbol in the staff), which typically indicate the duration and tone, respectively. In the case of Mensural notation, this duality is more general because even symbols that do not denote any sound (such as *rests*) may also appear at different heights, which is useful for reading the music when many rests appear consecutively.

Our work aims at producing transcripts that are useful, amongst other applications, for preservation purposes. Consequently, we are mainly interested in *diplomatic transcription*, which contain all the details needed to convey information about how the notation was written in the source. Thus, our system is designed to always recognize both the specific shape and height of each symbol, even in the





















Group	Symbol			
	Note	Semibrevis	Minima	Col. Minima
				
Rest	Longa	Brevis	Semibrevis	Semiminima
				
Clef	C Clef	G Clef	F Clef (I)	F Clef (II)
				
Time	Major	Minor	Common	Cut
				
Others	Flat	Sharp	Dot	Custos
				

Table 1: Representative elements of Mensural notation. These elementary symbols are depicted without considering their height or vertical position on the staff.

cases where the musical meaning is equivalent. For example, recognizing a *minima* as a *coloured minima* or failing to detect the position of a *rest* will be considered errors, although they do not modify the meaning of the music itself. These issues will be taken into account in the evaluation metrics adopted in our experiments.

3.2 Corpus

As a case of study, we consider the *Capitan* dataset presented in [8]. It contains a complete 96–page manuscript of the 17th century corresponding to a *missa* (sacred music), for which the only ground-truth available consists of diplomatic transcripts, without any symbol-to-image alignments or any other symbol segmentation information. For this corpus, a recognition baseline was defined in [8], which allows us to easily evaluate and compare the improvements attained by the proposed approach.

Each page contains 6 staves (music pentagram lines), some of which may be empty of content. The segmentation of page image into staff-section images is straightforward and methods such as that proposed in [4] can be generally used for almost perfect results. This simple stave segmentation step was carried out once for all pages when the dataset was created [8].

3.2.1 Image pre-processing

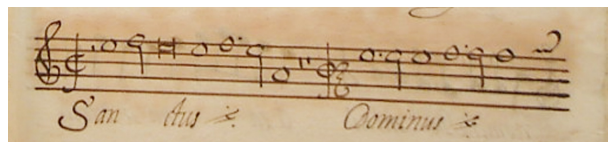
Staff-section images need to be pre-processed so that they are presented to the system in a convenient way.

A complete example of our pre-processing pipeline is shown in Fig. 1. It encompasses the following successive steps for the staff section depicted in Fig. 1a:

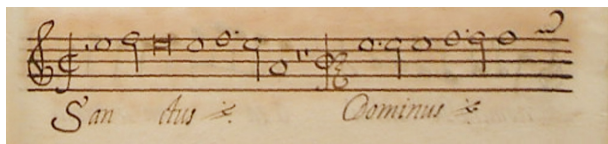
1. Skew correction (Fig 1b): the image skew is computed and corrected so that the staff remains aligned with the horizontal axis. We use the staff-line detection algorithm proposed in [9] which capitalizes on the excellent reference provided by the staff lines themselves.
2. Staff location (Fig 1c): to ensure that the staff section to be processed is well framed, we force the central line of the staff to be in the center of the image. In addition, the image is cropped so that it has a fixed height of 1.5 times the distance between the first and last line of the staff.
3. Height normalization: the recognition methods to be applied require that each image column (often referred to as “*frame*”) be of a fixed height. Therefore, the image is rescaled to a fixed height, without changing the aspect ratio.
4. Feature extraction (Fig 1d): finally, the image is represented as a set of meaningful features. With the hope of emphasizing the information needed for the HMR task, each frame is represented in three different ways: mean gray-scale values, horizontal gradient, and vertical gradient. This feature extraction method has been successfully used in similar tasks such as Handwritten Text Recognition [26].

As a result, each staff sample is finally represented as a variable-length sequence $\mathbf{x} = \vec{x}_1, \vec{x}_2, \dots, \vec{x}_T$ of d -dimensional feature vectors, where T is (propor-

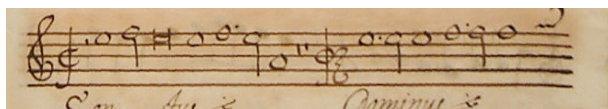
tional to) the staff image width. The actual value of d is adjusted experimentally.



(a) Original staff sample.



(b) Skew correction.



(c) Staff location and normalization.



(d) Feature extraction.

Fig. 1: Pre-processing steps (b-c) applied to convert an original staff image (a) into a feature vector sequences (d), represented also as an “image”.

Note that neither the staff section separation nor this pre-processing removes the accompanying text (lyrics), which is just considered “noise” for music notation recognition.

4 Framework

Let a staff-section image be represented as a sequence \mathbf{x} of feature vectors and let $\mathbf{s} = s_1 \dots s_m$, $s_j \in \Sigma, 1 \leq i \leq m$, be a sequence of musical symbols. We look for a most likely sequence of musical symbols $\hat{\mathbf{s}}$ according to:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} P(\mathbf{s} | \mathbf{x}) = \arg \max_{\mathbf{s}} p(\mathbf{x} | \mathbf{s}) P(\mathbf{s}) \quad (1)$$

where the factor $1/p(\mathbf{x})$ has been ignored because it is equal for all possible sequences, \mathbf{s} .

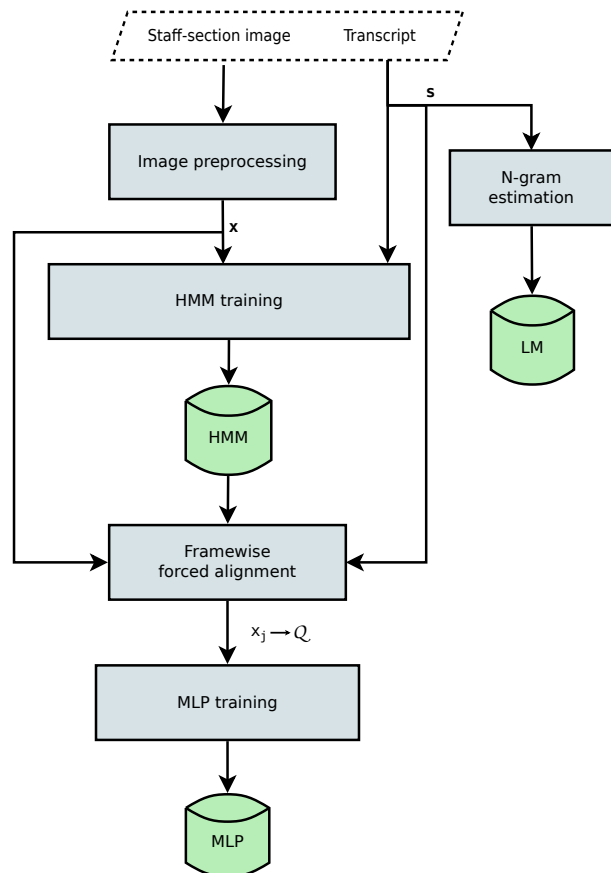


Fig. 2: An overview of the pipeline needed for the learning stage in the proposed framework for HMR.

In a basic formulation, the conditional density $p(\mathbf{x} | \mathbf{s})$ is typically approximated by means of Gaussian mixture HMMs. However, here HMMs are combined with a Multi-Layer Perceptron (MLP), which allows better discriminative behavior without losing the ability of HMMs to conveniently deal with variable-length sequences in a holistic way. This also allows the prior $P(\mathbf{s})$ to be properly formulated along with the MLP-HMMs, as in classical HMMs. In particular, the use of N -grams is considered, which increases the recognition accuracy thanks to *a priori* information about symbol concatenation likelihoods.

For the sake of clarity, the whole workflow for using HMM with discriminative training based on MLP is illustrated in Fig. 2 and Fig.3, for the learning and the decoding stages, respectively. The following subsections delve into the steps necessary to develop such framework.

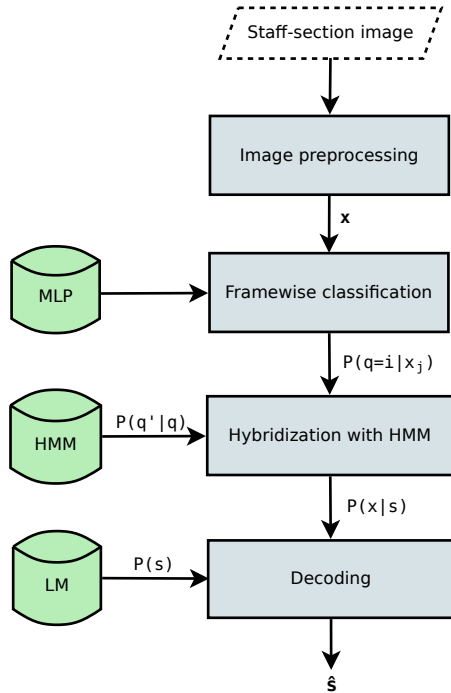


Fig. 3: An overview of the pipeline needed for the decoding stage in the proposed framework for HMR. The required models are those generated during the learning stage (see Fig.2).

4.1 Hidden Markov Models

This sub-section reviews the work presented in [7], where classical Gaussian mixture HMMs were proposed to model $p(\mathbf{x} | \mathbf{s})$. This review is necessary as a basis for the hybrid MLP-HMM proposed here.

Let $X \subseteq \mathbb{R}^D$ be a D -dimensional space of observations, where D is typically the number of features per frame, d , used to represent the input images. A continuous-density HMM [23] is a finite state machine which models the stochastic generation of sequences of vectors from X . It is defined by a finite set of states $Q \subset \mathbb{N}$, special initial and final states $I, F \notin Q$, a first-order state transition probability distribution $P(q' | q)$, $q' \in Q \cup \{I\}$, $q \in Q \cup \{F\}$, $P(I | F) \stackrel{\text{def}}{=} 0$, and a state-conditional observation emission distribution, $p(\vec{x} | q)$, $\vec{x} \in X$, $q \in Q$.

In order to approximate $p(\mathbf{x} | \mathbf{s})$ in Eq. (1), each music symbol s_j of \mathbf{s} is modelled by an HMM with Gaussian Mixture Model (GMM) state-conditional emission density. It is important to remark that each s_j is defined by both the shape and height of the corresponding music symbol. Therefore, it is necessary to consider a different HMM for each possible combination of these two aspects. If we assume \mathcal{Q} be

the the sets of *all* the (adequately relabelled) states of all these HMMs, $p(\mathbf{x} | \mathbf{s})$ is computed as:

$$p(\mathbf{x} | \mathbf{s}) = \sum_{\mathbf{q} \in \mathcal{Q}^*} p(\mathbf{x}, \mathbf{q} | \mathbf{s}) \\ = \sum_{\mathbf{q} \in \mathcal{S}(\mathbf{s}, T)} \prod_{i=1}^T P(q_i | q_{i-1}) p(\vec{x}_i | q_i) \quad (2)$$

where \mathcal{Q}^* is the set of all finite-length sequences of HMM states and $\mathcal{S}(\mathbf{s}, T)$ is the set of state sequences of length T corresponding to the concatenation of m music-symbol HMMs which model $\mathbf{s} = s_1 \dots s_m$.

Fig. 4 illustrates an HMM for the musical symbol “semibrevis-4s” $\in \Sigma$ (semibrevis in the fourth vertical space) modelling two “semibrevis-4s” notes in a staff fragment of Fig. 1a, represented as a sequence of feature vectors, as in Fig. 1d.

4.1.1 Maximum-Likelihood GMM-HMM training

Both the number of states per symbol (and therefore the overall number of states in \mathcal{Q}) and the number of Gaussians per state must be adjusted empirically [14]. Once these “HMM topologies” have been set for all symbols, the corresponding state-initial, state-transition and GMM distribution parameters can be trained from whole, unsegmented staff-section images, accompanied by the corresponding transcripts into sequences of musical symbols.

Maximum-Likelihood (ML) estimation of these parameters is easily carried out using a well-known instance of the Expectation-Maximization algorithm called *forward-backward* or Baum-Welch re-estimation [23]. This approach was first proposed and empirically evaluated for Mensural-notation HMR in [7].

4.1.2 Discriminative GMM-HMM training

ML estimation assumes a generative optimization, which can limit the accuracy of the trained models for recognition tasks such as HMR. Unlike ML, Discriminative Training (DT) explicitly aims at optimizing the model capacity to discriminate among competing classes (music symbols in our case).

The use of DT to train Gaussian-mixture HMMs for Mensural-notation HMR was first proposed in [8], where an estimation criterion based on the so called “*Minimum Phone Error*” (MPE) was adopted¹. MPE

¹ This term comes from the Speech Recognition community. Here, “*phones*” refer to the music symbols.

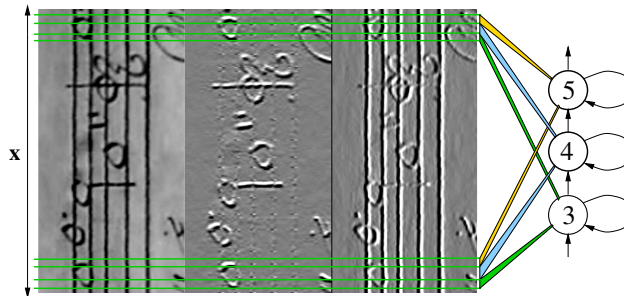


Fig. 4: HMM modelling of two symbol notes “semibrevis-4s” for the feature vector sequence of Fig. 1d.

focuses on minimizing the Levenshtein distance between the hypotheses and the music symbol sequences of the correct transcripts.

As discussed in [8], the optimization makes use of Symbol Lattices. For each training sample, two HMM-symbol-marked lattices are produced: the first one represents the correct symbol sequence and the second accounts for competing hypotheses. Using these lattices, HMM parameters are iteratively optimized according to the MPE criterion by means of a modified version of the Extended Baum-Welch re-estimation algorithm [21].

4.2 Discriminative training with Neural Networks

The state-conditional distribution $p(\vec{x}|q)$ can be approximated using other models, instead of GMMs. To better understand this generality, consider the following question: which is the most likely hidden state that may have emitted a given vector \vec{x} of the input sequence? This question corresponds to a classification task where the classes are states from \mathcal{Q} . Therefore, any statistically interpretable classifier explicitly or implicitly computes $p(q|\vec{x}) \forall q \in \mathcal{Q}$ which (as discussed later on) can be easily adapted to model $p(\vec{x}|q)$.

A simple way to train such a generic classifier is to first align each training sequence vector with a corresponding HMM state. This can be done by means of a “framewise forced alignment” [28], using HMM models, trained as discussed in Sec. 4.1. Then, the classifier training problem becomes a conventional supervised classification task, where the inputs are feature vectors from X and the outputs are corresponding hidden states in \mathcal{Q} .

As mentioned above, in this work we adopt a neural network discriminative classifier; namely a deep MLP which basically consists of a dense neural network organized in several layers. The last layer en-

compasses $|\mathcal{Q}|$ binary (*one-vs.-rest*) classifiers with *softmax* activation functions. These classifiers suit especially well in the context of our statistical framework because output activations can be properly interpreted as posterior probabilities [5]. More specifically, for each input vector \vec{x}_i of a feature vector sequence \mathbf{x} modelling a staff section, the trained MLP outputs an estimate of $P(q|\vec{x}_i), \forall q \in \mathcal{Q}$. A graphical illustration of this setup is shown in Fig. 5.

The number of input units of the MLP is the number of feature vector components, D . In a straightforward implementation, this value would be equal to d , *i.e.*, the number of features extracted from each frame of the input stave section image (see Sec. 3.2.1). Nevertheless, the overall optical discriminating power can be enhanced by adding to each frame a number of adjacent contextual frames. Therefore, the actual configuration of the MLP consists of an input layer of $D = (2c + 1)d$ neurons, where d is the number of features per frame and c is the length of the context considered. This layer is followed by a series of hidden layers that finally connect with the output layer of $|\mathcal{Q}|$ neurons. An empirical analysis of how many hidden layers are most adequate for the proposed task will be carried out in Sec. 5.3.

MLPs can be straightforwardly trained by gradient-descent using the well-known *back-propagation* algorithm. It should be pointed out, however, that as the number of layers increases (*i.e.*, the network becomes “*deep*”), conventional gradient descent needs to be assisted by deep Neural Network techniques such as layer pre-initialization, stochastic gradient descent, rectifiers, or learning rate scheduling [13]. See details in Sec. 5.3.

4.3 Statistical Language Models

As any other language, music notation exhibits regularities that, despite being extremely difficult to

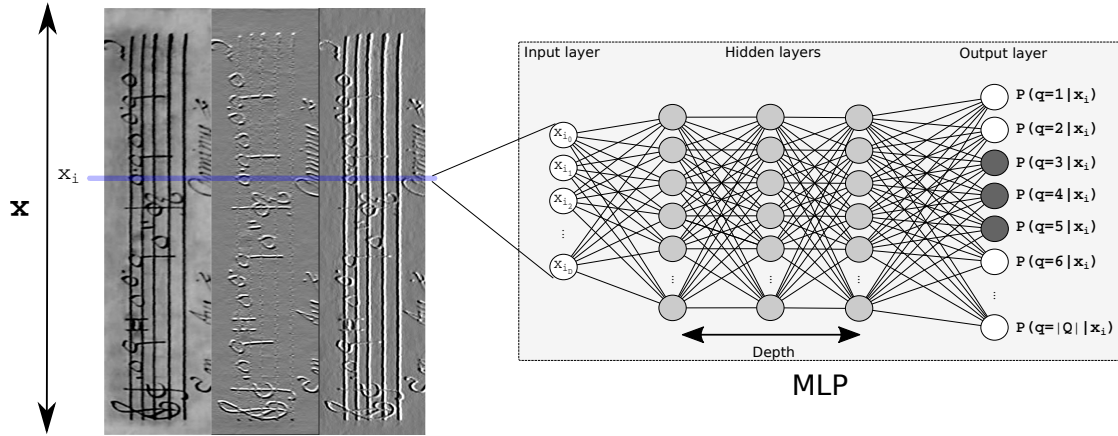


Fig. 5: The MLP is trained to estimate the posterior probability of each HMM state, given a frame of the input staff section. The output neurons that correspond to the states of the HMM that models “semibrevis-4s” (see Fig. 4) are highlighted.

model in their totality, can be exploited to some extent to improve the accuracy of the recognition. In this work, we resort to N -gram models to estimate the prior probability $P(\mathbf{s})$ needed in Eq. (1). An N -gram model assumes a local-context simplification of the probability of a sequence $\mathbf{s} = s_1 \dots s_m$ as²:

$$\begin{aligned} P(\mathbf{s}) &= P(s_1) \prod_{i=2}^m P(s_i | s_1 \dots s_{i-1}) \\ &\approx \prod_{i=1}^m P(s_i | s_{i-N+1} \dots s_{i-1}) \end{aligned} \quad (3)$$

where $P(s_i | s_{i-N+1} \dots s_{i-1})$ denotes the probability of finding s_i after $s_{i-N+1} \dots s_{i-1}$. These probabilities are the parameters of the N -gram model, which are easily estimated using training staff image transcripts [29].

Given the limited amount of data and the vocabulary considered, many events might not appear in the training set. In order to generalize better, a Knesser-Ney smoothing strategy [17] is considered so that no sequence has a non-zero probability.

4.4 Decoding

Once all the components have been adequately trained, a optimization or “decoding” process is carried out to compute Eq. (1); *i.e.*, to provide a best symbol sequence hypothesis $\hat{\mathbf{s}}$, given an input staff section

² For the sake of notation simplicity, for any sequence \mathbf{z} if $j < 1$, $P(z_k | z_j \dots z_{k-1})$ is assumed to denote $P(z_k | z_1 \dots z_{k-1})$. If $j = 1$, it is just $P(z_1 | \lambda) \equiv P(z_1)$, where λ is the empty sequence.

image represented as a sequence of T feature vectors $\mathbf{x} = \vec{x}_1, \vec{x}_2, \dots, \vec{x}_T$. Using HMMs to model $p(\mathbf{x} | \mathbf{s})$, as discussed in Sec 4.1, Eq. (1) becomes:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} P(\mathbf{s}) \sum_{\mathbf{q} \in \mathcal{S}(\mathbf{s}, T)} \prod_{i=1}^T P(q_i | q_{i-1}) p(\vec{x}_i | q_i) \quad (4)$$

In order to make the decoding process feasible, the sum is approximated by the dominating addend:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} P(\mathbf{s}) \max_{\mathbf{q} \in \mathcal{S}(\mathbf{s}, T)} \prod_{i=1}^T P(q_i | q_{i-1}) p(\vec{x}_i | q_i) \quad (5)$$

In the hybrid MLP-HMM approach, the state-conditional emission probabilities $p(\vec{x}_i | q_i)$ can be easily derived from the MLP output state posteriors, $P(q_i | \vec{x}_i)$, as:

$$p(\vec{x}_i | q_i) = \frac{P(q_i | \vec{x}_i)}{P(q_i)} p(\vec{x}_i) \quad (6)$$

and Eq. (5) becomes:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} P(\mathbf{s}) \max_{\mathbf{q} \in \mathcal{S}(\mathbf{s}, T)} \prod_{i=1}^T P(q_i | q_{i-1}) \frac{P(q_i | \vec{x}_i)}{P(q_i)} \quad (7)$$

where the factor $\prod_{i=1}^T p(\vec{x}_i)$ has been ignored because it is equal for all music symbol sequences (only depends on the input staff-section image). The hidden state priors $p(q_i)$ can be straightforwardly estimated from state frequencies of occurrence in the forced alignments used to train the MLP (*c.f.*, Sec. 4.2).

Equation (7) can be solved using the Viterbi algorithm [16]. Thus, given an input sequence of feature vectors \mathbf{x} , an output sequence of recognized musical symbols $\hat{\mathbf{s}}$ is obtained.

In practice, some additional factors must be taken into account. First, optical HMM probability densities ($p(\mathbf{x} | \mathbf{s})$ in Eq. (1) or the corresponding products of Eq. (7)) tend to be small and may become negligible as compared with the much higher values of language model probabilities, $P(\mathbf{s})$. A Grammar Scale Factor (GSF) is typically used to scale (the logarithm of) $P(\mathbf{s})$ in order to achieve an adequate balance between these two kinds of models. Second, the so-called Word Insertion Penalty (WIP) is used to weight the transition between output sequence symbols, in order to control the tendency of the decoder to produce shorter or longer sequences. Both the GSF and the WIP must be tuned empirically.

5 Experiments

Previous empirical work carried out with GMM-HMMs is reviewed and compared with the results of new experiments using the hybrid MLP-HMM approach here proposed. For this approach, we first analyze how different MLP configuration parameters affect the recognition performance. Next, we compute final results and compare them with previous works on the same corpus. Finally, we study the relationship between the size of the training set and the performance of the different models considered for HMR.

5.1 Corpus and assessment measures

The corpus considered here, referred to as *Capitan* is described in detail in [8], where a standard partition into training, validation, and test samples was established. A summary of the characteristics of the dataset as regards to this partition is given in Table 2.

	Training	Validation	Test
Staves	462	57	57
Different symbols	176	123	115
Running symbols	10 323	1 286	1 254

Table 2: Partition of the Capitan dataset, reporting the number of staves, the number of different music symbols (or “vocabulary”) and the number of running symbols.

Taking into account the different elements of the HMR task, we have considered several metrics to measure the recognition performance; namely:

- *Diplomatic Symbol Error Rate* (SER): computed as the average number of elementary editing operations needed to produce a reference (correctly transcribed) symbol sequence from the recognized symbol sequence.
- *Glyph Error Rate* (GER): as in SER but only taking into account the shape of the symbols, ignoring the height component (where any).
- *Height Error Rate* (HER): as in SER but only taking into account the height of the symbol. Those symbols that have no height are grouped into the same one.

5.2 HMM Setup

All the GMM-HMM results reported below correspond to the experiments carried out in [8], where configuration parameters were adjusted on the Capitan validation set in order to boost the HMR performance.

A strict left-to-right topology without loops was adopted for all symbol-level HMMs, with a variable number of states depending on the graphical width of the music symbol. This configuration resulted in a total of 178 symbol-level HMMs (176 music symbols plus 2 auxiliary characters), with a total of 3664 hidden states. The number of Gaussian functions in the emission mixtures was set to 4 and input frames were characterized by vectors of 180 dimensions, that is, 60 components per type of graphical feature (cf. Sec. 3.2.1).

GSF and WIP parameters are also optimized on the validation set for each language model considered.

The very same configuration is used here for the forced alignment process required for MLP training in the new MLP-HMM experiments.

All the experimentation has been conducted using the HTK toolkit [31], including the hybridization with MLP. Since HTK can only decode directly using up to 2-gram models, experiments with higher-order N -grams have been carried out by applying the re-scoring method [31, 19].

5.3 Configuration of the Neural Network

We start analyzing the impact of basic MLP configuration parameters on the overall HMR accuracy. In

particular, we evaluate two hyper-parameters: the number of context frames used to form each MLP input vector — which determines the size of the input layer — and the number of hidden layers (depth of the network). Some other components are fixed in advance so that the number of hyper-parameters does not become huge: the size of all hidden layers is established as $\frac{1}{2}io$, where, i and o are respectively the sizes of the input and output layers. As previously discussed, $i = 180(2c + 1)$ and $o = 3664$.

The number of hidden layers to consider ranges from 1 to 5. We make use of a series of Neural Network techniques that allow training such deep MLPs. Thus, the activation function for all the neurons is the so called *Rectified Linear Unit*, except for those of the output layer which use a *softmax* activation. Network weights are layer-wise pre-initialized according to a uniform distribution. The training itself is carried out by means of stochastic gradient descent, with a mini-batch size of 40 samples. The loss is measured with a categorical cross-entropy function. The learning rate is initially fixed to 0.002, although we use the NEWBOB scheduler which modifies the learning rate after each epoch according to the performance over the validation set [31].

Table 3 shows the SER results — as a general measure — on the validation set. As it can be observed, the results vary up to more than 3 percentage points depending on the specific configuration. The results confirm that using no frame context ($c = 0$) is less informative (27.06 of SER at best). It may seem that increasing the context is always beneficial, as long as the depth of the network is also increased. However, this reaches a limit quickly because a context of 5 frames, with a 23.87 of SER in the best case, fails to outperform the best value using a context of 3 frames, with 23.72 of SER. The set of values considered for the number of layers of the network seems to be representative enough, given that a local minimum is found in each column of Table 3. Nevertheless, it is possible that the amount of data available is hindering the possible benefits of using a wider frame context.

The best configuration corresponds to a context of 3 frames ($D = 1260$) and 2 hidden layers. This shall be adopted in the remaining experiments.

5.4 Comparison with previous works

In this section the accuracy of the approach here proposed is compared with previous work using GMM-

Hidden layers	Context length (c)			
	0	1	3	5
1	27.60	25.43	25.74	26.44
2	28.08	25.12	23.72	25.58
3	27.06	25.12	24.57	25.19
4	27.68	25.51	25.54	23.87
5	27.84	25.43	25.35	24.49

Table 3: Symbol error rate (SER, in %) with respect to increasing frame context and MLP depth over the validation set. Best values per column are typeset in boldface.

HMMs, both trained with the classic Maximum Likelihood (ML) estimation [7] and with Discriminative Training [8]. Further to these works, we have made experiments similar to those of the previous section on the same data partition to study the impact of increasing the frame context on GMM-HMM accuracy. However, results have not shown significant improvements. Therefore, we consider no context ($c = 0$) for GMM-HMMs, and the set-up proposed in the aforementioned references is kept.

Figure 6 shows the test-set SER results obtained with the different HMR models considered, with N -gram LMs up to $N = 4$. In all the cases, all the meta-parameters involved were optimized using the validation set.

The main limitation of GMM-HMMs stems from their discriminative ability. Even when they are trained in a discriminative way, SER improves from 52.7% to 46.4% (without language model). However, the use of MLP is able to very significantly boost the accuracy, further reducing the error rate down to 30.6% (also without language model). This corresponds to a 34% relative improvement over the best results so far.

The impact of the language model is evident in all the cases. While music generally constitutes a language of enormous complexity, it seems that there are (rather simple, albeit significant) regularities that can be easily captured computationally. It is clear that this information is of great help to improve the recognition of the music notation. At best, the use of 3-grams allows decreasing the error from 30.6% to 25.7% (a 16% relative improvement). The 4-gram models are probably under-trained with the limited data available and therefore they lead to a slightly worse performance. Similar tendencies are also observed for the other approaches, which substantiates the complementary nature of the language model with respect to the optical component of HMR.

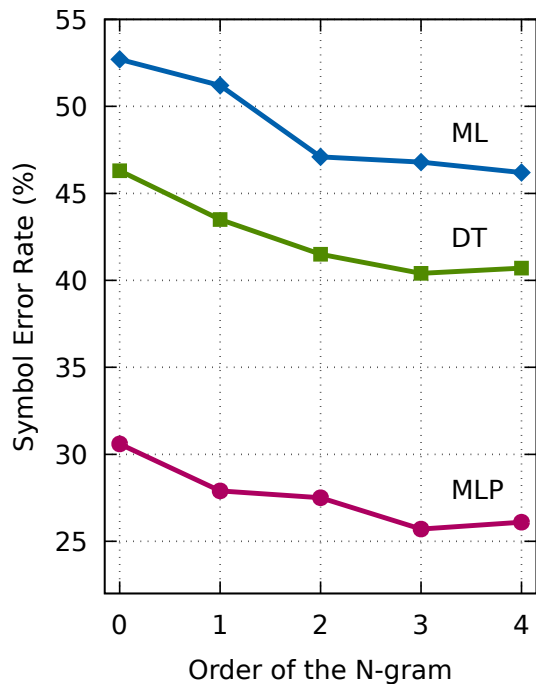


Fig. 6: SER (in %) obtained by increasing the order of N -gram LMs over the test set. Three HMR approaches are compared: GMM-HMMs, with both ML estimation [7] and DT [8], and the here proposed MLP-HMM (considering the best configuration according to Table 3).

Table 4 shows the final results considering all the aforementioned metrics. For the sake of comparison, we also report the performance of some previous research on automatic recognition of Early music notation, namely Aruspix system [22] and the segmentation-based approach described in [6]. In both cases, the performance of the recognition is quite poor. Aruspix misclassifies almost all symbols, thus yielding an error close to 100%. The segmentation-based approach does correctly recognize more symbols, provided they are correctly isolated. However, the process produces an over-segmentation — many actual symbols are broken into more than one isolated component — and so the final result is poor as well. It is really important to emphasize that this evaluation is not totally fair, since these methods were not designed to work with handwritten notation, which is the object of study here. However, including them might help to understand the need for holistic approaches based on machine learning, like the one proposed in this work, for handwritten notation. We can see that holistic approaches based on HMM, regardless of their specific configuration

(ML, DT or MLP), achieve significantly better results. Within this context, as reported above, the combination of HMM with MLP represents the approach that ends up obtaining the best recognition results by improving its discriminative capacity with specialized classifiers.

As regards to the dual nature of the symbols, it follows from GER and HER figures that both the height and the shape symbol components contribute almost equally to the accuracy. The HER is systematically lower in all cases, which is not surprising considering that there are 35 different shapes but only 16 different height positions.

Approach	SER	GER	HER
Aruspix [22]	94.5	93.0	94.1
Segmentation based [6]	93.3	91.7	90.4
HMM-ML [7]	46.2	41.2	34.9
HMM-DT [8]	40.4	35.2	28.2
HMM-MLP	25.7	22.4	18.7

Table 4: Summary of final results for different approaches. Note that the two first approaches (*Aruspix* and *segmentation based*) were not designed to deal with handwritten notation (more elaboration about this is found in the main text).

5.5 Analysis upon the size of the training set

The previous section has shown that the replacement of Gaussian emissions with the outputs of a MLP leads to important accuracy improvements. However, in the context of the HMR task, it is interesting to develop models that not only work well but that the amount of data necessary to train them is not very demanding. The handwriting style of Mensural manuscripts is very heterogeneous, given the large number of copyists at that time. It is likely that specific training data is needed for each type of writing style and, therefore, it is important that the recognition models work with a limited ground-truth size that can be obtained with little cost. Thus, in this section we study the accuracy achieved by all the HMM-based approaches proposed so far (ML, DT, and MLP) with respect to the size of the training set.

Figure 7 shows the results of such experiment. Results are given in terms of SER considering the best N -gram LM of each case ($N=4$ for ML, and

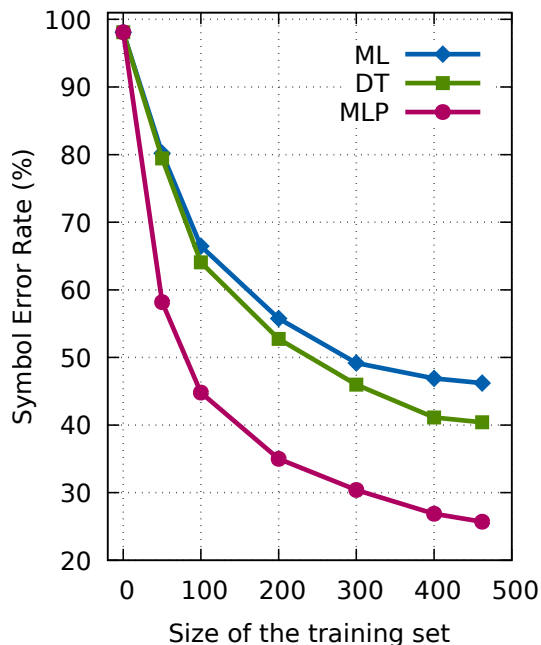


Fig. 7: SER achieved by the different HMM-based schemes with respect to the number of staff samples in the training set. Results are given considering their best N -gram LM ($N=4$ for ML, and $N=3$ for DT and MLP).

$N=3$ for DT and MLP). The most striking remark that can be extracted from these results is that the use of MLP not only ends up producing significantly better results with a reasonably-sized dataset, but the superiority over other schemes also holds for few training samples. Moreover, the best result obtained by previous approaches with all available data is outperformed by the MLP approach with just 200 training samples, and it achieves a SER below 30% with only 300 training staff sections.

Thus, the use of MLP not only performs significantly better than previous approaches but it also uses the available training data in a more profitable way. It can be observed that the use of DT also represents an improvement with respect to the traditional (ML) HMM training approach but to a much lesser extent.

6 Conclusions

In this work, a previously unexplored approach for the recognition of Handwritten Music in Mensural notation has been presented. It relies on the use of Hidden Markov Models whose emission probabilities

are derived from the outputs of a Multi-Layer Perceptron, trained on pairs of training images and their corresponding transcripts. The goal is to improve the discriminant capabilities of traditional GMM-HMMs. This hybrid model is combined with N -gram statistical language models, trained on the training transcripts. As in the case of classical GMM-HMMs, here N -grams also help biasing the recognizer towards the a-priori most likely sequences, aiming to further increase the overall recognition accuracy.

The experiments carried out have shown that the approach studied in this work attains an error rate around 25% at symbol level. This result greatly overcomes the performance achieved in previous works that reported error rates around 40%, at best. The new results definitely evince that real transcription applications are clearly feasible using the proposed methods.

Although the present work represents a considerable improvement compared to previous approaches, there is still room for further advances for which several avenues for future research are worth exploring. For instance, the double shape-height nature of symbols is one of the main features that distinguish music notation from other similar domains such as text. In this work we considered that each combination of these two types of elements must be understood by the system as a totally independent symbol. Nevertheless, all notes of the same shape share many features, and this information is not being used to improve recognition performance. Also, once language models have proven to contribute noticeably to improve recognition accuracy, a more convenient estimation can possibly be obtained from independent shape and height statistics.

Furthermore, given the limited amount of data and the relatively large number of symbols, data augmentation represents an interesting framework to consider. However, it should be noted that data augmentation for musical staves must go beyond simple blurring, rotation, and scaling. In particular, staff lines are elements that should remain basically similar in all images, and only musical symbols should be altered.

Compliance with Ethical Standards

Funding: *Hidden because of double-blind review.*

Conflict of Interest: Authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Bainbridge, D., Bell, T.: The challenge of optical music recognition. *Computers and the Humanities* **35**(2), 95–121 (2001)
- Bertolami, R., Bunke, H.: Hidden markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition* **41**(11), 3452–3460 (2008)
- Bluche, T.: Deep neural networks for large vocabulary handwritten text recognition. Ph.D. thesis, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (2015). Discipline : Informatique
- Bosch, V., Calvo-Zaragoza, J., Toselli, A.H., Vidal-Ruiz, E.: Sheet music statistical layout analysis. In: 15th International Conference on Frontiers in Handwriting Recognition, ICFHR 2016, Shenzhen, China, October 23–26, 2016, pp. 313–8 (2016)
- Bourlard, H., Wellekens, C.: Links between markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(11), 1167–1178 (1990)
- Calvo-Zaragoza, J., Barbancho, I., Tardón, L.J., Barbancho, A.M.: Avoiding staff removal stage in optical music recognition: application to scores written in white mensural notation. *Pattern Analysis and Applications* **18**(4), 933–943 (2015)
- Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Early handwritten music recognition with hidden markov models. In: 15th International Conference on Frontiers in Handwriting Recognition, ICFHR 2016, Shenzhen, China, October 23–26, 2016, pp. 319–324 (2016)
- Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten music recognition for mensural notation: Formulation, data and baseline results. In: 14th International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, August 13–15, 2017, pp. 1081–1086 (2017)
- Cardoso, J.S., Capela, A., Rebelo, A., Guedes, C., Pinto, J.: Staff detection with stable paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(6), 1134–1139 (2009)
- Espana-Boquera, S., Castro-Bleda, M.J., Gorbemoya, J., Zamora-Martinez, F.: Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(4), 767–779 (2011)
- Fujinaga, I., Hankinson, A., Cumming, J.E.: Introduction to SIMSSA (single interface for music score searching and analysis). In: Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLfM@JCDL 2014, London, United Kingdom, September 12, 2014, pp. 1–3 (2014)
- Gallego, A., Calvo-Zaragoza, J.: Staff-line removal with selectional auto-encoders. *Expert Syst. Appl.* **89**, 138–148 (2017)
- Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. The MIT Press (2016)
- Günter, S., Bunke, H.: HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and gaussian components. *Pattern Recognition* **37**(10), 2069–2079 (2004)
- Hankinson, A., Burgoyne, J.A., Vigliensoni, G., Fujinaga, I.: Creating a large-scale searchable digital collection from printed music materials. In: Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16–20, 2012 (Companion Volume), pp. 903–908 (2012)
- Jelinek, F.: *Statistical methods for speech recognition*. MIT Press (1998)
- Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: 1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08–12, 1995, pp. 181–184 (1995)
- Lee, S., Son, S.J., Oh, J., Kwak, N.: Handwritten music symbol classification using deep convolutional neural networks. In: Information Science and Security (ICISS), 2016 International Conference on, pp. 1–5. IEEE (2016)
- Ortmanns, S., Ney, H., Aubert, X.: A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language* **11**(1), 43–72 (1997)
- O'Shaughnessy, D.: Automatic speech recognition: History, methods and challenges. *Pattern Recognition* **41**(10), 2965–2979 (2008)
- Povey, D.: *Discriminative Training for Large Vocabulary Speech Recognition*. Ph.D. thesis, University of Cambridge (2003)
- Pugin, L.: Optical music recognition of early typographic prints using hidden markov models. In: ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8–12 October 2006, Proceedings, pp. 53–56 (2006)
- Rabiner, L., Juang, B.H.: *Fundamentals of speech recognition*. Prentice hall (1993)
- Ramirez, C., Ohya, J.: Automatic recognition of square notation symbols in western plainchant manuscripts. *Journal of New Music Research* **43**(4), 390–399 (2014)
- Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marçal, A.R.S., Guedes, C., Cardoso, J.S.: Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval* **1**(3), 173–190 (2012)
- Toselli, A.H., Juan, A., Vidal, E.: Spontaneous handwriting recognition and classification. In: 17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23–26, 2004., pp. 433–436 (2004)
- Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal interactive transcription of text images. *Pattern Recognition* **43**(5), 1814–1825 (2010)
- Toselli, A.H., Romero, V., Vidal, E.: Alignment between text images and their transcripts for handwritten documents. *Language Technology for Cultural Heritage* pp. 23–37 (2011)

29. Vidal, E., Thollard, F., De La Higuera, C., Casacuberta, F., Carrasco, R.C.: Probabilistic finite-state machines-part ii. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(7), 1026–1039 (2005)
30. Yang, W., Tao, J., Ye, Z.: Continuous sign language recognition using level building based on fast hidden markov model. *Pattern Recognition Letters* **78**, 28–35 (2016)
31. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al.: *The HTK book*, vol. 3.5. Cambridge: Entropic Cambridge Research Laboratory (2015)