UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dep. de Sistemes Informàtics i Computació

Representació de patrons musicals a partir de codificació mitjançant sub-seqüències freqüents de notes

Treball Fi de Màster

Màster Universitari en Intel·ligència Artificial, Reconeixement de Formes i Imatge Digital

AUTOR/A: Menárguez Box, Aitana

Tutor/a: Vidal Ruiz, Enrique

Cotutor/a: Toselli, Alejandro Héctor

Director/a Experimental: Villarreal Ruiz, Manuel

CURS ACADÈMIC: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

TREBALL FI DE MÀSTER

Representació de patrons musicals a partir de codificació mitjançant sub-seqüències freqüents de notes

Autora:
Menárguez Box, Aitana

Tutor:
Vidal Ruiz, Enrique
Co-tutor:
Toselli, Alejandro Héctor
Director experimental:
Villarreal Ruiz, Manuel

Treball final del Màster Universitari en Intel·ligència Artificial, Reconeixement de Formes i Imatge Digital*del*
Departament de sistemes informàtics i computació (DSIC)

realitzat al

Pattern Recognition and Human Language Technology (PRHLT) Research Center

Curs 2023-24

Resum

Aquest treball estudia la representació de patrons musicals mitjançant sub-seqüències freqüents de notes. Entre altres objectius, es tracta de millorar el rendiment d'un sistema de reconeixement de partitures històriques manuscrites. Per assolir-ho, es proposa l'ús de la tecnologia de codificació BPE (*Byte Pair Encoding*). Aquest enfocament pretén explorar unitats semàntiques musicals que permetisquen millorar el comportament del sistema actual, que fins ara ha treballat amb unitats musicals independents (notes separades).
A llarg termini, s'aspira a experimentar amb unitats de significat musical que transcendisquen la notació simbòlica per tal de millorar l'accessibilitat i comprensió dels repertoris musicals en entorns digitals, així com la recuperació de la informació dins d'aquestos. Aquest treball representa un esforç per abordar els reptes actuals en la representació, reconeixement i cerca d'informació musical en manuscrits musicals antics, amb implicacions per a diverses aplicacions al camp de la música i la tecnologia.

Resumen

Este trabajo estudia la representación de patrones musicales mediante la codificación de sub-secuencias frecuentes de notas. Entre otros objetivos, se trata de mejorar el rendimiento de un sistema de reconocimiento de partituras manuscritas históricas. Para lograrlo, se propone el uso de la tecnología de codificación BPE (*Byte Pair Encoding*). Este enfoque busca explorar unidades semánticas musicales que permitan mejorar el comportamiento del sistema actual, que hasta ahora ha trabajado con unidades musicales independientes (notas separadas).
A largo plazo, se aspira a experimentar con unidades de significado musical que trasciendan la notación simbólica con el propósito de mejorar la accesibilidad y comprensión de los repertorios musicales en entornos digitales, así como la recuperación de la información dentro de estos. Este trabajo representa un esfuerzo por abordar los desafíos actuales en la representación, reconocimiento y búsqueda de información musical dentro de manuscritos musicales antiguos, con implicaciones para diversas aplicaciones en el campo de la música y la tecnología.

Abstract

This work studies the representation of musical patterns by encoding frequent subsequences of notes. Among other objectives, it aims to improve the performance of a system for recognizing historical handwritten sheet musics. To achieve this, it proposes the use of the BPE (Byte Pair Encoding) technology. This approach seeks to explore musical semantic units that allow improving the behavior of the current system, which until now has worked with independent musical units (individual notes).
In the long term, it aims to experiment with units of musical meaning that transcend symbolic notation with the aim of improving the accessibility and understanding of musical repertoires in digital environments, as well as the retrieval of information within them. This work represents an effort to address current challenges in the representation, recognition and search of musical information inside ancient musical manuscripts, with implications for various applications in the fields of music and technology.

Agraïments

La realització d'aquest treball fi de màster ha sigut possible gràcies a l'ajuda de ValgrAI, *Valencian Graduate School and Research Network of Artifficial Intelligence*.

Índex

Índex de figures

Índex de taules

Índex d'algoritmes

Capítol 1

Introducció

La preservació i digitalització de textos manuscrits antics ha sigut una preocupació central en el camp de les humanitats digitals durant les últimes dècades. La necessitat d'aquesta tasca es fa encara més evident en el cas dels manuscrits musicals, on la informació continguda en partitures històriques és de gran valor en l'àmbit d'estudi de la musicologia i la història de la música.

Els manuscrits litúrgics representen especialment una font important per a entendre la pràctica musical d'èpoques passades, però la seua conservació es troba en risc constant a causa del deteriorament dels materials originals i la dificultat de transcripció manual. Així, la digitalització d'aquestes partitures no només facilita la seua conservació sinó que també possibilita noves vies d'investigació a través de l'anàlisi digital.

El reconeixement de text manuscrit ha aconseguit grans avanços en els últims anys gràcies a l'ús de xarxes neuronals profundes, millorant significativament el procés de transcripció automàtica de documents històrics. No obstant això, el camp del reconeixement de partitures manuscrites presenta reptes addicionals.

Mentre que en els textos escrits el model lingüístic subjacent pot estar ben definit, amb estructures sintàctiques i semàntiques pròpies de les llengües naturals, en la música aquesta estructura lingüística no és tan clarament definida. Tradicionalment, els sistemes de reconeixement automàtic de música manuscrita han tractat de processar els documents nota a nota, sense considerar la possibilitat d'agrupar els símbols musicals en unitats significatives, com es fa en l'àmbit de reconeixement de text, ja que el concepte de paraula no es troba dins la música.

El treball detallat en aquesta memòria pretén superar l'anomenada limitació a partir d'introduir el concepte de *paraula musical*, amb l'ús del *byte pair encoding* (BPE) per generar agrupacions de símbols que aporten informació contextual addicional als models òptics de reconeixement de text musical. L'objectiu és explorar com la utilització de BPE pot ajudar a millorar els resultats del reconeixement òptic de les partitures.

D'altra banda, altre dels majors reptes que es presenten en aquest context és la quantitat limitada de dades disponibles per a l'entrenament dels models. La transcripció manual de manuscrits musicals antics és una tasca costosa i en molts casos no és possible comptar amb un volum de dades transcrites prou gran. Per aquest motiu, el present treball també explora com afecta l'ús del BPE, enriquint el procés d'aprenentatge amb informació musical més àmplia, en el rendiment dels sistemes

de reconeixement amb conjunts de dades reduïts.

1.1 Estructura de la memòria

Aquesta memòria està organitzada per proporcionar una visió integral del treball realitzat i facilitar la comprensió dels diversos aspectes abordats. En primer lloc, al capítol 2, es presenta una revisió de l'estat de l'art que aborda les aplicacions actuals del reconeixement òptic de música, així com les limitacions dels enfocaments existents. També es detallen els aspectes clau del sistema de reconeixement adoptat, incloent l'entrenament i el modelat estadístic del llenguatge.

El capítol 3 es dedica al plantejament del problema, es descriuen el sistema actual i les dades d'entrenament, així com la solució proposada basada en BPE. Es detallen les característiques de la codificació BPE i es planteja la base teòrica sobre la qual es desenvoluparà la solució. La descripció del conjunt de dades ocupa capítol 4, on s'ofereix informació general sobre les dades utilitzades, el seu format i les particions experimentals aplicades.

En el capítol número 5 es detalla el marc experimental del treball, incloent proves preliminars realitzades amb BPE i el detall de les configuracions d'entrenament dels models òptics. Més tard, el capítol 6 s'enfoca en les mesures d'avaluació emprades i els resultats obtinguts junt amb una discussió sobre aquestos. També s'analitzen les implicacions dels resultats i es revisen les codificacions BPE utilitzades.

Finalment, la memòria conclou amb el capítol 7 on es discuteixen les limitacions del treball, les conclusions extretes i les possibles direccions de futurs estudis. Aquesta secció proporciona una perspectiva final sobre els avenços aconseguits i les àrees que podrien ser explorades en investigacions posteriors.

Capítol 2

Estat de l'art

Abans d'aprofundir en l'explicació del treball realitzat, a aquest capítol es farà una revisió del marc teòric al qual s'engloba el projecte.

Per una banda, s'esmentaran de forma superficial les tecnologies existents, pel que fa al nostre coneixement del camp, en matèria de reconeixement d'elements musicals en partitures manuscrites antigues. D'altra banda, s'aprofundirà en l'explicació de les tècniques que concerneixen i s'han emprat dins d'aquest treball fi de màster.

2.1 Reconeixement òptic de música

El reconeixement òptic de música (OMR[1]) és l'aplicació del reconeixement òptic de caràcters (OCR[2]) per tal d'interpretar de forma automàtica les partitures en un format editable i/o reproduïble [41]. A partir d'una imatge musical, manuscrita o impresa, l'OMR ens permetria modificar, escoltar i inclús cercar (entre altres [16]) el contingut dins de la partitura. Quan l'OMR s'aplica a partitures manuscrites, s'anomena reconeixement manuscrit de música (HMR[3]).

Al llarg del temps, l'OMR s'ha referenciat també com a OCR musical, malgrat hi ha diferència entre els dos. Mentre que l'OCR musical tindria com a objectiu el reconeixement únicament dels símbols (com a unidimensionals), l'OMR també se centra en reconèixer la complexitat dins d'una partitura: seqüències ordenades vertical i horitzontalment, relacions espacials, notacions rítmiques, notacions harmòniques, diferents veus i instruments... [32]

El terme "reconeixement òptic de música" és més bé genèric perquè cada tasca englobada dins d'aquest depèn de diferents factors com per exemple el tipus de notació de les partitures (occidental moderna, mensural, neumàtica...) o el tipus de gravat de les mateixes (manuscrit o imprès).

Existeixen sistemes que aborden l'OMR com a tasques involucrades dins d'un procés més gran a partir de la segmentació de símbols [39], beneficiant-se que la notació musical no té entitats de baix nivell com els fonemes o els caràcters.

No obstant això, la segmentació de símbols en música manuscrita antiga és particularment difícil a causa del soroll, la degradació dels documents i la baixa qualitat

[1] De l'anglès *Optical Music Recognition*.
[2] De l'anglès *Optical Character Recognition*.
[3] De l'anglès *Handwritten Music Recognition*.

de les imatges. Alguns mètodes han reformulat les etapes inicials d'aquest procés com a tasques de detecció d'objectes, a partir de l'ús de xarxes neuronals profundes basades en regions [6, 7, 26]. Dins de [8] es proporciona una base per detecció directa d'objectes musicals a partitures experimentant amb diferents models i dades de diverses topologies.

D'altra banda, altres investigacions [23] van aplicar un enfocament holístic[4], a partir de models de Markov ocults (HMMs[5]), a on el model du a terme un reconeixement complet: partint d'una imatge inicial, s'obté finalment una seqüència musical. Posteriorment, es van millorar amb tècniques d'entrenament discriminatives i models híbrids. Tot i que els HMM són adequats per a aquesta tasca, no han arribat al mateix nivell d'avanç que dins dels camps del reconeixement de veu o text manuscrit, que s'han beneficiat de la utilització de xarxes neuronals profundes més recentment.

Així, dins d'aquest treball fi de màster, com es perfilarà més endavant, s'ha abordat el problema plantejat a partir de l'ús de xarxes neuronals profundes per al reconeixement holístic *end-to-end* de manuscrits musicals antics, seguint la línia de treball de [3].

2.1.1 Aplicacions del reconeixement de manuscrits musicals (HMR) antics

Quan es parla d'HMR aplicat a partitures manuscrites antigues, com és el cas d'estudi dins d'aquest treball, hi ha diversos factors que contribueixen a augmentar la complexitat de la tasca: diferències de notació entre manuscrits, variacions estilístiques, qualitat dels manuscrits... Al llarg dels anys, s'han intentat abordar aquestos problemes a partir de diferents enfocaments, centrant-se, principalment, en la digitalització de les partitures i la implementació de sistemes per realitzar tasques de cerca sobre elles.

Un dels projectes més rellevants en aquest àmbit ha estat Cantus Ultimus[6], enfocat en reconeixement òptic de manuscrits medievals. Aquesta ferramenta permet cercar en llibres de cants, emprant contorns melòdics relatius i mètodes tradicionals d'OMR de recuperació d'informació. Malgrat això, al igual que ocorre en altres propostes com Liber Usualis[7] o OMMR4All [28], la cerca es realitza a partir de les hipòtesis de major probabilitat generades per a cadascun dels símbols (notes) reconeguts. Aquest enfocament *1-best* genera els resultats més precisos que els anomenats sistemes poden aconseguir, però pot ignorar, en la majoria dels casos, informació rellevant per a la recuperació d'informació que milloraria la cerca si es consideraren probabilísticament totes les hipòtesis possibles.

Per altra banda, també s'han implementat sistemes de cerca en col·leccions digitals més modernes. Un exemple és el projecte de Musiconn ScoreSearch[8]. Aquest

[4] Holístic en aquest cas fa referència a que el procés és global, que no aborda el problema per parts ni seqüencialment. La segmentació està inclosa dins del procés i es tracta de forma general el reconeixement.

[5] De l'anglès *Hidden Markov Models*

[6] Ferramenta accessible a través de `https://cantus.simssa.ca/`.

[7] Ferramenta accessible a través de `https://liber.simssa.ca/`.

[8] Ferramenta accessible a través de `https://www.musiconn.de/services/musiconnscoresearch`.

sistema permet realitzar una cerca transportada, és a dir, es poden buscar fragments de partitures sense importar la tonalitat original del manuscrit, la qual cosa permet obtindre més informació encara sobre els manuscrits. Aquestos avenços són notables, és clar, però continuen sense abordar el problema complet de la cerca en partitures manuscrites a partir d'un sistema probabilístic més robust.

2.1.2 Limitacions dels enfocaments actuals

A pesar dels avenços en el camp del reconeixement i cerca en manuscrits antics, els enfocaments actuals tenen limitacions clares. La cerca basada en hipòtesis *úniques* descarta l'aprofitament d'informació addicional que podria millorar els resultats. Ja dins del reconeixement de text manuscrit (HTR[9]) s'ha demostrat que l'ús d'indexació probabilística millora significativament la qualitat de la recuperació d'informació assolida [35] . Encara que a [5] s'ha estudiat aquest enfocament aplicat a partitures manuscrites antigues, encara no s'ha explotat aquest plenament per a l'HMR.

2.2 Detalls del sistema d'HMR adoptat

Com s'ha esmentat anteriorment, dins d'aquest treball s'adopta la mateixa estratègia que a [3] per a la tasca de reconeixement de partitures manuscrites antigues.

La principal característica d'aquest sistema és l'ús de xarxes neuronals convolucionals recurrents (CRNNs[10]) per modelar la probabilitat a posteriori de generar uns símbols determinats d'eixida donada una imatge d'entrada. S'assumeix que les imatges d'entrada són regions aïllades (tetragrames) dins de la partitura que han sigut detectades prèviament a partir de tècniques ximples, conegudes i robustes com [29].

Una CRNN està composta per un bloc de capes *convolucionals* seguit per altre bloc de capes *recurrents* [33]. A cada capa convolucional, normalment, li segueix altra de *max-pooling* que redueix la dimensionalitat de l'eixida. El bloc convolucional s'encarrega d'extreure característiques rellevants de la imatge i les capes recurrents interpreten aquestes característiques en termes de seqüències d'eixida de símbols musicals. A aquest treball, les xarxes utilitzades estan compostes per unitats anomenades "Long Short Term Memory" (LSTM) organitzades en una arquitectura LSTM bidireccional (BLSTM) [12]. A més, gràcies a les capes del bloc convolucional no és necessari cap procés manual d'extracció de característiques, ja que aquestes són entrenades de forma automàtica a partir de dades específiques per extreure implícitament les millors característiques per a aquesta tasca concreta [43].

Les unitats d'activació in l'última capa convolucional poden ser vistes com una seqüència de vectors de característiques que representen la imatge d'entrada x o també com a versions linealment reduïdes de x. Sent W l'amplada de x, l'amplada de les imatges de característiques resultats serà $J = \gamma W$, a on $\gamma \leq 1$ ve definida pels paràmetres de *max-pooling*.

[9] De l'anglès *Handwritten Text Recognition*.
[10] De l'anglès Convolutional Recurrent Neural Networks

El bloc convolucional produeix tantes imatges de característiques com a quantitat de filtres que s'hagen establert a l'última capa. Després, totes aquestes imatges es concatenen per formar una única imatge de característiques que és introduïda a la primera capa de BLSTM. Les unitats d'activació de l'última capa recurrent, llavors, són considerades estimacions de les probabilitats a posteriori per finestra de la imatge:

$$P(\sigma|x,j), \quad 1 \leq j \leq J, \quad \sigma \in \Sigma' \stackrel{\text{def}}{=} \Sigma \cup \epsilon \tag{2.1}$$

a on Σ és un conjunt de símbols musicals i ϵ és un símbol especial de "no caràcter" usat en imatges que contenen dues o més instàncies consecutives del mateix símbol musical [12]. $P(\sigma|x,j)$ sovint és anomenat el *posteriorgrama* de x.

2.2.1 Entrenament de la CRNN

Les xarxes neuronals convolucionals poden ser entrenades directament a partir del descens de gradient emprant el conegut algoritme de *Back Propagation* (BP). Les xarxes BLSTM es poden entrenar de forma similar amb una versió del BP anomenada *Back Propagation Through Time* (BPTT) [42].

L'algoritme convencional de BPTT necessita informació sobre quin símbol ha de ser predit en cadascuna de les finestres d'eixida però un conjunt d'entrenament convencional d'HMR només proporciona, per a cada imatge de regió de la partitura, la seua transcripció de símbols musicals corresponent, sense cap informació explícita de la localització espacial dels símbols. Oportunament, les capes BLSTM poden ser entrenades sense aquest tipus d'informació a partir de la funció de pèrdua *Connectionist Temporal Classification* (CTC) [13].

El procés resultant d'entrenament CTC és una forma de "expectació maximització" (EM) similar a l'algorimte *backward-forward* emprat per entrenar els HMMs [25]. A més, se sol afirmar que, per a un entrenament CTC adequat és necessari l'ús del símbol especial de "no caràcter" [13].

D'altra banda, per tal d'evitar *overfitting*, s'aplica *Dropout* [34] durant la iteració de descens de gradient, desactivant un conjunt d'unitats arbitràriament seleccionades. En aquest cas, solament s'ha aplicat a les unitats del bloc recurrent.

2.2.2 Modelat estadístic del llenguatge

La notació musical presenta regularitats o restriccions que, malgrat resultar extremadament complexe modelar-les plenament, poden ser aprofitades fins a cert punt per tal de millorar la precisió de reconeixement [3]. Aquestes regularitats són anomenades, normalment, restriccions del llenguatge.

La CRNN modela de forma implícita aquestes restriccions, ja que ha sigut entrenada per estimar el *posteriorgrama* de símbol $P(\sigma|x,j)$, el qual es troba estretament relacionat, com es mostra més endavant, amb la probabilitat a posteriori $P(s|x)$ a on $s \in \Sigma^*$ és una transcripció d'una imatge determinada en una seqüència de símbols musicals. Anteriorment, en altres dominis relacionats amb HMR com a reconeixement de text manuscrit o reconeixement automàtic de veu (ASR[11]) s'ha demostrat

[11] De l'anglès *Automatic Speech Recognition*.

clarament que l'ús explícit de models de llenguatge (LM[12]) entrenats independentment pot millorar significativament la precisió de reconeixement [1].

Llavors, com a [3], s'han emprat models de llenguatge, en concret models d'N-grames per a la tasca d'HMR d'aquest treball. Un model d'N-grames assumeix una simplificació de context local per a la probabilitat d'una seqüència $s = s_1 \dots s_m$ com a[13]:

$$P(s) = P(s_1) \prod_{i=2}^{m} P(s_i|s1 \dots s_{i-1}) \approx \prod i = 1^m P(s_i|s_{i-N+1} \dots s_{i-1}) \tag{2.2}$$

a on $P(s_i|s_{i-N+1} \dots s_{i-1}$ denota la probabilitat de trobar s_i després de $s_{i-N+1} \dots s_{i-1}$. Aquestes probabilitats són els paràmetres del model d'N-grames i poden ser fàcilment estimades emprant conjunts de transcripcions d'entrenament [38].

Donada la limitació de la quantitat de dades d'entrenament, per tal de generalitzar millor, s'empra l'estratègia de suavitzat proposada per Kneser i Ney [18]. Així, $P(s) > 0 \ \ \forall s \in \Sigma^*$.

2.2.3 Reconeixement o decodificació

Formalment, la imatge d'entrada x ha de ser reconeguda o decodificada en una seqüència de símbols musicals més probable $\hat{s} \in \Sigma^*$:

$$\hat{s} = \arg\max_{s \in \Sigma^*} P(s|x) \tag{2.3}$$

L'equació 2.3, sense cap model de llenguatge, pot ser resolta a través d'optimització local. Per a aquest propòsit, es computa un símbol òptim per a cada posició del *posteriorgrama* j:

$$\hat{\sigma}_j = \arg\max_{\sigma \in \Sigma'} P(\sigma|x, j), \quad 1 \leq j \leq J \tag{2.4}$$

Llavors, es pot obtindre una eixida òptima aproximada amb:

$$\hat{s} = \hat{s}_1 \dots \hat{s}_m \approx \mathcal{F}(\hat{\sigma}_1 \dots \hat{\sigma}_J), \quad m \leq J \tag{2.5}$$

a on $\mathcal{F} : \Sigma'^J \rightarrow \Sigma^m$ és una funció que fusiona tots els caràcters consecutius com $\hat{\sigma}_j = \hat{\sigma}_{j-1}$ i després elimina tots els "no caràcters" $\sigma_j = \epsilon$ [13].

Per tal d'emprar un model de llenguatge, primer és necessari reescriure l'equació 2.3 com a $\hat{s} = \arg\max_{s \in \Sigma^*} P(s)\ p(x|s)$, a on $P(s)$ és la probabilitat del LM, calculada com a l'equació 2.2.
Si ara $\sigma = \sigma_1 \dots \sigma_J \in \Sigma'^*$ i assumint que x és condicionalment independent de s donat σ, $p(x|s)$ pot ser reescrit com a:

$$p(x|x) = \sum_{\sigma} p(x, \sigma|s) = \sum_{\sigma} P(\sigma|s)\ p(x|\sigma, s) = \sum_{\sigma} P(\sigma|s)\ p(x|\sigma) \tag{2.6}$$

[12] De l'anglès *Language Models*.

[13] Per simplificar la notació, per a qualsevol seqüència z si $j \leq 1$ $P(z_k|z_j \dots z_{k-1})$ s'assumeix que denota $P(z_k|z_1 \dots z_{k-1})$. Si $j = 1$, és simplement $P(z_1|\lambda) \equiv P(z_1)$ a on λ és la seqüència buida.

Es pot considerar $P(\sigma|s)$ com uniforme per a tota σ que complisca $s = \mathcal{F}(\sigma)$ i nul per a qualsevol altra σ. Aleshores,

$$p(x|s) \propto \sum_{\sigma:\mathcal{F}(\sigma)=s} p(x|\sigma) \approx \max_{\sigma:\mathcal{F}(\sigma)=s} p(x|\sigma) \tag{2.7}$$

Finalment, de les equacions 2.3 i 2.7 es té:

$$\hat{s} = \arg\max_{s} P(s)\, p(x|s) \approx \arg\max_{s} P(s) \max_{\sigma:\mathcal{F}(\sigma)=s} p(x|\sigma) \tag{2.8}$$

$p(x|\sigma)$ s'obté a partir del *posteriorgrama* de x de la següent manera:

$$p(x|\sigma) \approx p(x) \prod_{j=1}^{J} \frac{P(\sigma_j|x,j)}{P(\sigma_j)^k} \tag{2.9}$$

a on el factor $p(x)$ pot ser ignorat en l'equació 2.8 ja que no depèn de σ. les probabilitats a priori dels símbols $P(\sigma)$, $\sigma \in \Sigma'$ poden ser estimades directament a partir de les dades d'entrenament i k es tracta d'un meta-paràmetre que s'ajusta empíricament [1].
Clarament, l'equació 2.8 ja no pot ser resolta a partir d'optimització local, encara que es poden obtindre solucions suficientment precises emprant l'algoritme de Viterbi.

D'altra banda, el model d'N-grames es representa com un transductor d'estats finits a on els pesos dels vèrtex són el producte de les probabilitats d'N-grames (els factors de l'equació 2.2 i les probabilitats del model òptic (els factors de l'equació 2.9 sense el terme $p(x)$).

Finalment, l'equació 2.8 es resol utilitzant el decodificador cerca de feixos (*beam search*) de Viterbi implementat en la ferramenta de Kaldi [22].

La figura 2.1 il·lustra la canonada (*pipeline*) sencera per decodificar una imatge d'entrada x en una seqüència aproximadament òptima de símbols musicals $\hat{s}$ a partir d'una CRNN i un LM.

2.3 Implementació i ús del sistema adoptat: *PyLaia*

PyLaia és una ferramenta *open-source* d'aprenentatge profund (*deep learning*) dissenyada específicament per tasques seqüència a seqüència (*sequence-to-sequence*), amb èmfasi particular en reconeixement de text manuscrit (HTR).

Construït sobre *PyTorch*, *PyLaia* fa ús de tècniques de *deep learning* per abordar problemes complexos on les seqüències d'entrada i eixida poden tindre una longitud diferent. Per exemple, com és el cas d'aquest treball, el reconeixement de partitures manuscrites (HMR).

Aquesta ferramenta és el resultat de la implementació del marc teòric detallat en la secció anterior (vore figura 2.1).

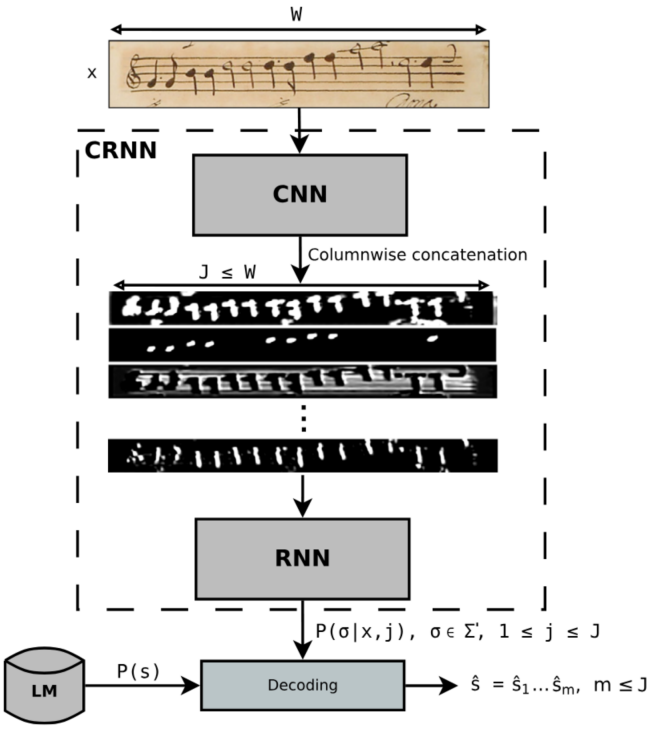


FIGURA 2.1: Visió general del sistema d'HMR emprat en aquest treball, des d'una imatge d'una regió musical d'entrada fins a la seua decodificació en una seqüència de símbols musicals a partir d'una CRNN i un LM.
Font: [3].

2.3.1 Característiques i funcionament

PyLaia compta amb una infraestructura flexible i modular, la qual cosa permet a qui la utilitza personalitzar els models i fluxos de treball en base a les seues necessitats. Les funcionalitats principals es construeixen al voltant de les següents parts:

- **Modelatge *sequence-to-sequence***: Destaca en tasques a on l'alineament entre les seqüències d'entrada (imatges de les línies musicals) i les d'eixida (transcripcions d'aquestes línies) són desconegudes.
- **Pèrdua CTC**: *PyLaia* empra la pèrdua CTC a l'entrenament dels models. Aquesta va ser dissenyada per gestionar tasques dins les quals el *timing* (cadència, ritme) de la seqüència d'eixida, relativa a la d'entrada, és incert. Açò és un element clau a l'hora de reconèixer documents manuscrits, on el símbols poden trobar-se distorsionats o espaiats de forma irregular.
- **Interfície de línia de comandaments**: gràcies a la implementació de l'ús de la ferramenta a través de comandaments, ha sigut possible desplegar els entrenaments, validacions i proves de les diferents experimentacions de forma senzilla.

2.3.2 Contextualització dins del projecte

Dins d'aquest treball, *PyLaia* ha sigut útil per a l'entrenament de models òptics de reconeixement de símbols musicals dins d'un manuscrit musical litúrgic històric. La

possibilitat que ofereix la ferramenta de gestionar tasques *sequence-to-sequence*, junt amb la seua implementació de la pèrdua CTC, han permès abordar les irregularitats i complexitats de la notació musical manuscrita antiga.

D'altra banda, *PyLaia* ha facilitat experimentar amb diferents grandàries de les particions de dades, possibilitant, a més, l'anàlisi del comportament de la compressió BPE (*Byte Pair Encoding*) proposada per a aquest projecte i detallada més endavant.

En resum, *PyLaia* ha proporcionat una plataforma robusta, flexible i efectiva per tractar els reptes plantejats al treball.

2.4 title

Byte Pair Encoding (BPE) Existeixen diversos treballs enfocats en codificar la informació musical d'alguna forma semblant a paraules [17, 19, 20]. En aquest treball, però, som s'explica a la secció 3.3, un dels objectius principals d'aquest treball és experimentar amb la codificació BPE (que es detallarà àmpliament al llarg de la memòria) per tractar de millorar el rendiment del sistema base de reconeixement de partitures manuscrites explicat anteriorment. La proposta s'inspira en l'ús convenient d'aquesta tècnica dins d'altres àmbits relacionats amb el HMR, com el processament del llenguatge natural (NLP[14]) [14, 30] o inclús dins del mateix àmbit del reconeixement de text musical [9, 11]. Encara que, tant als exemples esmentats com a la resta de bibliografia existent (fins a on arriba el nostra coneixement de la matèria) no s'empra el BPE amb la mateixa finalitat que la que es planteja a aquest treball.

Ací, se segueix la formulació que ha demostrat funcionar millor dins de l'HTR: amb BPE es pretén modelar unitats més grans que els caràcters aïllats, és a dir "paraules musicals", per tal de millorar el rendiment dels models òptics de reconeixement [27]. Tot açò s'explica detalladament al capítol 3.

[14] De l'anglès *Natural Language Processing*.

Capítol 3

Plantejament del problema i proposta de solució

Dins d'aquest capítol s'exposa en profunditat la problemàtica a abordar amb aquest treball: millorar un sistema de reconeixement de partitures manuscrites musicals antigues mitjançant la codificació de sub-seqüències freqüents de notes; així com una visió concreta i detallada de la solució proposada per abordar les qüestions endavant esmentades.

3.1 Context i justificació

3.1.1 El sistema actual

El sistema actual per al reconeixement de partitures manuscrites històriques plantejat a la secció 2.2, punt des del qual parteix aquest treball fi de màster, ha demostrat un rendiment raonablement bo [3] però presenta, encara, àrees susceptibles de millora. El sistema es basa en l'ús d'una CRNN amb entrenament CTC que s'entrena amb transcripcions musicals *símbol a símbol*. És a dir, transcripcions nota a nota del conjunt de dades que s'empre a l'aprenentatge.

Al haver utilitzat notes separades com a transcripcions a l'entrenament, les decodificacions seran, també, seqüències de notes separades. Això pot plantejar problemes, sobre tot a l'hora de recuperar informació dins de les partitures. Si es vol buscar una seqüència musical[1] dins d'un manuscrit reconegut pel sistema, s'haurà de fer nota a nota [4, 5]. Certament, aquest tipus de cerca suposa un cost temporal elevat, resultant prohibitiu quan és el cas de grans col·leccions de manuscrits.

Encara que l'aspecte de cerca musical no s'ha abordat directament dins d'aquest treball degut a la seua complexitat, les millores ací proposades s'enfoquen en ser útils eventualment per al procés de recuperació d'informació de seqüències musicals.

Així, doncs, a partir de codificar les dades a través de sub-seqüències freqüents de notes, s'espera que la cerca i recuperació d'informació musical als manuscrits, que s'explorarà en el futur, siga de millor qualitat i més eficient.

[1] Típicament, en contextos d'investigació musical, l'interès de la cerca radica en seqüències musicals (entre altres elements) i no només en notes aïllades. Al igual que per a textos manuscrits es busquen paraules i no lletres individuals. És per això que sempre que ací es parle de recuperar informació musical, es tractarà de seqüències i no símbols musicals.

3.1.2 Les dades d'entrenament

Altre dels reptes significatius, no només dins del reconeixement de partitures manuscrites antigues sinó en l'HTR i HMR en general, és la quantitat de dades d'entrenament que és necessària per obtindre resultats efectius. Com es comprovarà més endavant, en el cas d'estudi d'aquest treball la reducció en la quantitat de dades d'entrenament condueix a una disminució del rendiment del sistema.

Aquesta problemàtica resulta particularment rellevant, però, dins del context de les partitures manuscrites històriques en aplicacions a situacions reals, on les dades d'entrenament són costoses d'obtindre degut a l'esforç necessari per realitzar transcripcions a mà d'aquest tipus de documents.

Llavors, el plantejament que també s'explora en aquest treball és que la codificació proposada a partir de les sub-seqüències freqüents de notes, puga proporcionar informació addicional que cobrisca la mancança de dades i, per tant, millore el rendiment del model en situacions a les quals no es disposa d'una gran quantitat de dades d'entrenament.

3.2 La solució proposada

Per tal d'abordar les problemàtiques esmentades anteriorment, es pretén codificar les sub-seqüències freqüents de notes a partir de l'ús de *byte pair encoding* (vore més avall). Com s'ha apuntat a la secció 2.4, a través d'aquesta codificació s'aspira a "descobrir" unitats musicals més grans que els símbols (notes) aïllats naturalment provinents de les transcripcions d'entrenament. És a dir, es volen codificar *paraules musicals* per entrenar els models òptics, com se sol fer a HTR normalment, i, potencialment, millorar la qualitat del sistema d'HMR proposat.

Per tal de comprovar com aquesta proposta afecta al rendiment del sistema base, serà necessari entrenar i avaluar diferents models òptics amb dades codificades segons el marcs experimental i d'avaluació proposats als capítols 5 i 6.

3.2.1 El *Byte Pair Encoding* (BPE)

La codificació de parells de bytes o *byte pair encoding* [10] és una forma ximple de compressió de dades descrita per Philip Cage per primera vegada a l'any 1994. Aquesta tècnica converteix els bytes successius més recurrents en bytes "nous" creats artificialment.

Per exemple[2]: per a la cadena `aaabdaaabac` el *parell* de bytes literals (originals) que més vegades ocorre és `aa`, per tant se substitueix (en aquest cas d'esquerra a dreta) aquest parell per altre byte "artificial" que no s'haja utilitzat encara, com pot ser `Z`. Llavors, el procés de codificació per a la cadena donada es pot seguir de la següent forma: `aaabdaaabac` $\xrightarrow{\texttt{Z=aa}}$ `ZabdZabac` $\xrightarrow{\texttt{Y=ab}}$ `ZYdZYac`. Com que tots els parells de bytes originals que es troben a la nova seqüència ocorren només una vegada, l'algoritme podria acabar ací. Alternativament, el procés pot continuar amb *byte pair encoding* recursiu de forma que es reemplacen també els parells de bytes "artificials" i, per tant, l'última compressió possible seria `ZYdZYac` $\xrightarrow{\texttt{ZY=X}}$ `XdXac` ja que , a partir

[2]Exemple extret de https://en.wikipedia.org/wiki/Byte_pair_encoding.

d'aquest estat, a la cadena ja no hi queden parells (tant de bytes originals com "artificials") que ocórreguen més d'una vegada.

La descompressió de les dades s'obtindria fàcilment aplicant els reemplaçaments dels parells en ordre invers al que s'ha seguit anteriorment.

A partir dels reemplaçaments que es realitzen al llarg d'aquesta codificació, s'obté un *vocabulari* de parells. El pseudocodi d'aquest procés es pot vore més avall a l'algoritme 1. Ací no s'ha tingut en compte si els parells de bytes són o no originals (del vocabulari original); aquesta és l'aproximació que s'ha seguit al treball. Com

Algoritme 1 Pseudocodi de l'algoritme d'entrenament BPE

Requereix: $S = s_1 ... s_N$ ▷ Seqüència d'entrenament
  $N = |S|$ ▷ Talla d'S
  $\mathcal{V} = \cup_{i=1}^{N} \{s_i\}$ ▷ Vocabulari inicial
  for $i = 1 \ldots i = N$ **do** ▷ Comptabilitzar freqüències dels parells
    $F[s_i\ ,\ s_{i+1}] + +$ ▷ Si $F[s_i\ ,\ s_{i+1}]$ no existeix, es crea amb valor $= 0$
  end for
  $\hat{i} = \arg\max_{1 \leq i < N} F[s_i\ ,\ s_{i+1}]$ ▷ Trobar el parell més freqüent
  $p\ ,\ q = s_{\hat{i}}\ ,\ s_{\hat{i}+1}$
  while $F[p\ ,\ q] > 1$ **do** ▷ Mentres el parell tinga freqüència > 1
    for $i = 1\ ;\ i < N\ ;\ i + +$ **do** ▷ Substituir els parells per *tokens* a la seqüència
      if $s_i\ ,\ s_{i+1} == p\ ,\ q$ **then**
        $s_i = pq$
        $s_{i+1} = \lambda$ ▷ λ és la cadena buida
      end if
    end for
    $N = |S|\ ;\ \mathcal{V} = \cup_{i=1}^{N} \{s_i\}$ ▷ Actualitzar la talla d'S i el vocabulari
    for $i = 1 \ldots i = N$ **do** ▷ Actualitzar les freqüències dels parells
      $F[s_i\ ,\ s_{i+1}] + +$
    end for
    $\hat{i} = \arg\max_{1 \leq i < N} F[s_i\ ,\ s_{i+1}]$ ▷ Trobar el nou parell més freqüent
    $p\ ,\ q = s_{\hat{i}}\ ,\ s_{\hat{i}+1}$
  end while
  return $\mathcal{V}$ ▷ Torna el vocabulari après

s'observa, a partir de la seqüència d'entrada (d'entrenament), "s'aprèn" un vocabulari de *tokens* que es pot emprar per codificar qualsevol altra seqüència. El procés de codificació o *tokenització* a partir d'un vocabulari segueix l'esquema descrit a l'algoritme 2. Es tracta de codificar d'esquerra a dreta la seqüència d'entrada amb els *tokens* més llargs possibles.

D'aquesta manera, queden descrits dos processos a partir dels quals es podran entrenar diferents vocabularis BPE per realitzar diverses codificacions de les dades, com s'ha proposat abans en aquesta memòria.

3.2.2 Detalls d'implementació del BPE

A partir d'ara, per poder emprar la codificació BPE a l'entrenament dels models òptics, serà necessari codificar les dades disponibles. Per tant, s'aplicarà la codificació a les línies musicals de les transcripcions i es buscaran els parells consecutius més

Algoritme 2 Pseudocodi de l'algoritme de *tokenització* BPE

Requereix: $\mathcal{V}$; $S = s_1 ... s_N$ ▷ Vocabulari d'entrada i seqüència a codificar
  $N = |S|$ ▷ Talla d'S
  $S' = \lambda$ ▷ λ és la cadena buida
  $i = 1$; $f = N$ ▷ Inicialitzar posicions inicial i final
  while $i \leq N$ **do** ▷ Es pretén escanejar la seqüència d'esquerra a dreta
    $W = S_{i:f}$ ▷ W es una subseqüència d'S (paraula)
    if $W \in \mathcal{V}$ **then**
      $S'_{|S'|+1} = W$ ▷ S'afegeix la paraula al final d'S'
      $i = f$; $f = N$ ▷ S'actualitzen les posicions inicial i final
    else if $|W| == 1$ **then** ▷ Si és un símbol aïllat fora del vocabulari...
      $S'_{|S'|+1} = W$ ▷ ...es tokenitza com a símbol aïllat
      $i = f$; $f = N$ ▷ S'actualitzen les posicions inicial i final
    else
      $f - -$ ▷ Es redueix la paraula per la dreta
    end if
  end while
  return S' ▷ Torna la seqüència codificada

freqüents dins d'aquestes. És a dir, les línies actuaran com a seqüències que es processaran totes al mateix temps i els símbols musicals seran els bytes que, a partir d'ara, anomenarem *tokens*. Per tant, un *token* pot ser tant un byte original com un "artificial", ja que es buscaran les parelles de *tokens* consecutius indistintament del tipus que siga cadascun.

Agafant com a referència una implementació existent de l'algoritme BPE [31][3] s'ha dut a terme una pròpia per poder modificar i experimentar amb diversos paràmetres de codificació. En concret, s'han implementat dos hiperparàmetres distints per tal de construir vocabularis BPE amb granularitats diferents i permetre, per tant, realitzar múltiples entrenaments de models òptics (un per configuració):

MIN_OCC Mínim nombre d'ocurrències que ha de tindre un parell de *tokens* per compactar-se en un únic *token*. Si no es troba cap parell de *tokens* consecutius que complisca el mínim nombre d'ocurrències determinat, acabarà la codificació.

MAX_LEN Màxima longitud que es permet per a un *token*. Això és, màxima quantitat de símbols musicals originals que poden formar un *token*. Si, arribat a un punt, tots els potencials nous *tokens* tenen una longitud major que el paràmetre indicat (la quantitat de símbols originals que el conformen és major que MAX_LEN), acabarà la codificació.

Per exemple: suposant l'ús de minúscules per als *tokens* originals i l'ús de majúscules per als *tokens* "artificials", el parell de *tokens* $\{\texttt{X}, \texttt{Y}\}$ siguent $\texttt{X} = \texttt{aab}$ y $\texttt{Y} = \texttt{bb}$ no seria valid per a un MAX_LEN $= 4$, ja que el *token* resultant de la unió del parell seria $\texttt{Z} = \texttt{aabbb}$, que està format per 5 símbols originals.

D'altra banda, també s'ha implementat una llista de *tokens* especials per protegir-los de la codificació. Aquestos, en el cas d'estudi del treball, seran les claus musicals,

[3] A banda de l'article referenciat, existeix un repositori GitHub que pot ser accedit a través de https://github.com/soaxelbrooke/python-bpe/tree/master si es vol aprofundir en detall la implementació que ha sigut realitzada.

ja que les seqüències musicals que es volen trobar, han d'estar formades únicament per notes. Llavors, les claus seran *tokens* especials, amb un sol element, que no es tindran en compte a l'hora de cercar els parells de *tokens* consecutius més freqüents; hi haurà un per cada clau del manuscrit.

Notació real i exemple de codificació

Per clarificar la forma en què s'aplica aquesta codificació a les dades d'entrenament emprades a l'experimentació dins d'aquest treball, es mostra a continuació un exemple del procés de codificació per a un cas concret.

Suposant que es volen emprar per a l'entrenament les següents transcripcions de dues línies[4], seguint la notació exposada al capítol 4:

Línia 1: `f3 l2 l2 l2 s1 s0 s1 l2 l3 l2 l2`
Línia 2: `c4 l1 s0 l1 s1 s0 s1 l2 l2 s1`

Inicialment, els *tokens* dins de les línies són els símbols originals (notes i claus) i es troben separats per espais en blanc que s'ignoren a l'hora de realitzar la cerca de parells. Per *entrenar un BPE* (obtindre un vocabulari BPE) a partir de diverses línies (abans referides com a seqüències), els parells de *tokens* consecutius es busquen de forma conjunta en la totalitat de les línies. El vocabulari inicial seria $\{$`c4`, `f3`, `l1`, `l2`, `s0`, `s1`$\}$, amb una talla igual a 6.

Suposant ara els hiperparàmetres per al BPE de MIN_OCC $=$ 2 i MAX_LEN $=$ 2, el parell que més ocurrències té en aquest cas és `l2 l2` (apareix tres vegades en el conjunt de línies) i, llavors, s'ha de compactar. Per compactar dos *tokens*, s'usa un caràcter especial unificador qualsevol `#`[5]. És a dir que ara el parell `l2 l2` passarà a ser un únic *token*: `l2#l2`.

Línia 1: `f3 l2#l2 l2 s1 s0 s1 l2 l3 l2#l2`
Línia 2: `c4 l1 s0 l1 s1 s0 s1 l2#l2 s1`

A continuació, es troben dos parells amb el mateix nombre d'ocurrències (2): `s1 s0` i `s0 s1`. Quan hi ha ambigüitat per escollir quin parell compactar, s'elegix per ordre alfabètic. És a dir que tocaria compactar `s0 s1` com a `s0#s1`.

Línia 1: `f3 l2#l2 l2 s1 s0#s1 l2 l3 l2#l2`
Línia 2: `c4 l1 s0 l1 s1 s0#s1 l2#l2 s1`

Ara, l'únic parell que compleix la restricció del mínim nombre d'ocurrències per considerar-se un nou *token* és `s1 s0#s1`. Aquest parell, però, si es compactara, passaria a formar el *token* `s1#s0#s1`, que té una longitud de 3 (conté 3 símbols, notes, originals). Al no complir la restricció de màxima longitud permesa per als *tokens* i no quedar cap altre parell que complisca ambdues restriccions de l'algoritme alhora, ací terminaria la codificació. El vocabulari resultant s'extrauria de les línies codificades

[4] Es tracta d'un exemple clarificador ximple, per entrenar un model òptic és necessari una major quantitat de línies.

[5] El tipus de caràcter unificador, si és que es decideix emprar, no afecta als resultats de l'entrenament dels models. S'ha elegit aquesta notació només per claredat visual, però podrien unificar-se els *tokens* només ajuntant el parell sense un espai al mig (`1212`).

finals i seria {c4, f3, l1, l2, l2#l2, s0, s0#s1, s1}, amb una talla igual a 8.

3.2.3 Format de les dades d'entrenament codificades

Com és d'esperar, per a un mateix conjunt de línies d'entrenament, es poden realitzar diferents codificacions BPE (segons els hiperparàmetres escollits) i per comprovar quin és l'efecte d'aquestes en el rendiment del model òptic, és necessari entrenar un model per a cadascuna de les codificacions. Més endavant, al capítol 5, es realitza una experimentació preliminar per escollir un criteri de selecció dels millors valors d'hiperparàmetres de BPE per al cas d'estudi del manuscrit proposat.

En relació a la forma de les dades d'entrenament per als models òptics amb *paraules musicals* extretes a través de BPE, es planteja una qüestió: el format específic que han de tindre. Habitualment, quan es realitza l'entrenament amb transcripcions de símbols musicals individuals i com que s'empra una funció de pèrdua CTC (vore capítol 2), s'introdueixen els símbols de les línies al model intercalant-los amb un separador <space> de manera que la transcripció d'una línia qualsevol tindria la forma: <space> t_1 <space> ... <space> t_N <space>. Siguent t_n cadascun dels símbols musicals de la transcripció per a una línia determinada.

Una vegada s'ha realitzat la codificació BPE de les dades d'entrenament, però, no es tenen símbols musicals aïllats, sinó *tokens* musicals. La primer forma evident de formatejar aquestes dades BPE seria afegint un separador <space> de la mateixa manera que abans; aquesta vegada per separa els *tokens* musicals en compte dels símbols musicals aïllats: <space> c_1 <space> ... <space> c_N <space>. Siguent c_n cadascun dels *tokens* musicals de la transcripció codificada amb BPE per a una línia determinada. Un exemple concret, com és el cas de la línia 1 de la secció 3.2.2, tindria l'aspecte:

```
<space> f3 <space> l2#l2 <space> l2 <space> s1 <space> s0#s1 <space>
l2 <space> l3 <space> l2#l2 <space>
```

En el cas anterior, que anomenarem entrenament amb "*tokens* BPE unificats", els símbols que aprèn el model són els *tokens* del vocabulari entrenat per BPE. En el cas concret, aprendria els del vocabulari de talla 8 esmentat anteriorment. És a dir, les *paraules musicals* senceres.
També es pot plantejar el mode de formatejar de les dades d'entrenament d'altra manera: amb "*tokens* BPE dispersos". Així, el símbol unificador # passaria a ser un espai en blanc, de manera que la transcripció d'una línia qualsevol tindria la forma: <space> c_1^1 ... c_1^K <space> c_2^1 ... c_2^L <space> ... <space> c_N^1 ... c_N^M <space>. Siguent c_i^j el símbol original *i*èsim del *token* *j*èsim. L'exemple anterior, doncs, quedaria com segueix:

```
<space> f3 <space> l2 l2 <space> l2 <space> s1 <space> s0 s1 <space>
l2 <space> l3 <space> l2 l2 <space>
```

D'aquesta forma els símbols que aprendria el model serien els mateixos que per al cas original (sense emprar BPE): els símbols musicals aïllats. Per al cas concret de l'exemple anterior, aprendria el vocabulari de talla 6 esmentat. D'aquesta forma, el model haurà de predir els símbols original alhora que prediu la posició del separadors <space>. Aquesta aproximació és la més semblat a la usada per entrenar

models d'HTR habitualment.

En el capítol 5 també s'experimentarà amb ambdós tipus de format per a les dades d'entrenament BPE.

3.3 Objectius principals del treball

Per recapitular aquest capítol, a continuació es presenten els objectius principals que es pretenen assolir amb aquest treball:

- **Analitzar l'efecte del *byte pair encoding*:** observar com evoluciona el rendiment del sistema de reconeixement de text manuscrit musical plantejat amb les diferents modificacions proposades per al *byte pair encoding*, incloent entre elles el format de les dades una vegada hagen sigut codificades.
- **Explorar l'impacte de la mida del conjunt de dades d'entrenament:** avaluar com la reducció de dades d'entrenament afecta a la precisió en el reconeixement de símbols musicals.
- **Plantejar un marc de treball a futur:** en concret per la cerca de seqüències musicals dins de manuscrits antics.
- **Contribuir a la digitalització i preservació de manuscrits litúrgics antics:** aportar noves eines i metodologies que facilitin la transcripció automàtica i la recuperació d'informació d'aquest tipus de documents, assegurant-ne la seva preservació i accessibilitat futura.

Capítol 4

Descripció del conjunt de dades

L'experimentació realitzada aquest treball, d'entrenament de models òptics d'HMR a partir de diferents configuracions, s'ha dut a terme utilitzant el manuscrit musical referenciat com a *Cod. 253* [24]. Aquest pertany a la col·lecció de la biblioteca de l'abadia de Vorau i ha estat proporcionat per l'Acadèmia Austríaca de Ciències. Es tracta d'un document històric que conté exemples de notació musical antiga, en concret de notació germànic-gòtica, i es data al voltant de l'any 1450.

A la Fig. 4.1 es mostren exemples del manuscrit que, d'ara endavant, es referirà com a Vorau-253.

4.1 Informació general

El manuscrit es troba escrit predominantment amb tetragrames (quatre línies horitzontals) i alguns pentagrames (cinc línies horitzontals) [1]. La notació, és monofònica [2], és a dir que només hi ha una única línia melòdica en cadascuna de les pàgines. Tampoc hi ha cap tipus d'anotació rítmica, no s'inclouen indicacions sobre la durada de les notes. A banda de les línies musicals, el manuscrit conté text que acompanya les melodies per poder-les cantar, típicament per tractar-se de música d'origen litúrgic.

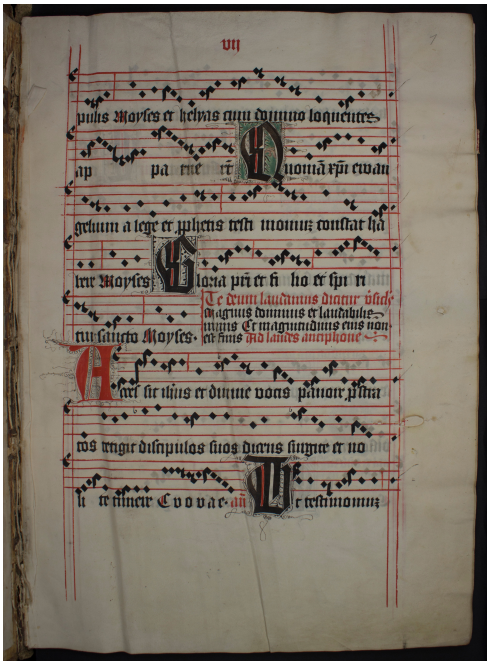
El Vorau-253 està compost per un total de 490 pàgines, incloent-hi la portada i la contraportada. Dins d'aquestes pàgines, es troben aproximadament mil tetragrames i un centenar de pentagrames.

Les pàgines del manuscrit han estat digitalitzades en imatges a color amb una resolució de 3247 píxels d'amplada i 4407 píxels d'altura. La informació associada a cadascuna de les imatges ha sigut estructurada en fitxers XML, per tal de facilitar el seu processament automàtic. Dins d'aquests fitxers es troben, entre altres dades, les transcripcions de les línies tant musicals com de text del manuscrit. Cadascuna d'aquestes ve acompanyada de l'anotació de les coordenades per a la seua segmentació. En total, al manuscrit hi ha anotades 6095 línies de text i 5055 línies musicals.

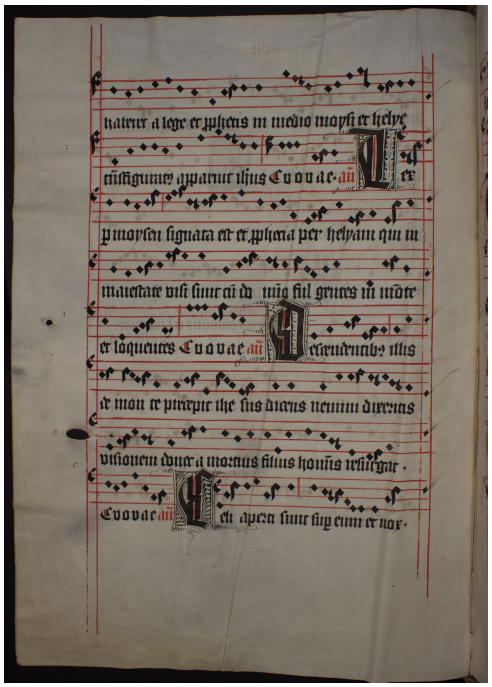
4.2 Format de les dades

Degut a la natura d'aquest projecte, s'ha descartat de l'estudi l'ús de les línies de text i només s'ha treballat amb les musicals. Així, doncs, el tipus de manuscrit considerat, admet una notació senzilla que en treballs anteriors [3 - 5] ha demostrat ser realment adequada per realitzar experiments bàsics d'entrenament i prova amb imatges dins

[1] Per tal de facilitar la lectura, al llarg d'aquesta memòria s'empraran els termes *tetragrama*, "línia musical", o simplement línia per referir-se tant a tetragrames com a pentagrames, ja que, en aquest cas que ens ocupa, no importa si les línies musicals estan formades per quatre o cinc línies horitzontals.



Pàgina *0001_A-VOR253-001r*



Pàgina *0002_A-VOR253-001v.jpg*

FIGURA 4.1: Exemples d'imatges de les pàgines del manuscript musical de Vorau-253. Font: [24].

d'aquest àmbit.

Des de la perspectiva del reconeixement òptic, una imatge d'una partitura pot ser concebuda com una seqüència vertical de línies horitzontals (els tetragrames) a on es "dibuixen" una sèrie de símbols musicals (les notes). Una nota pot localitzar-se tant en una línia com en un espai entre dues línies (o per sobre o sota una sola línia si es tracta d'un extrem del tetragrama). En el llenguatge musical, la posició d'una nota al tetragrama es fonamental per determinar la seua afinació, el so que representa. D'aquesta forma, per fer un ús adequat dels models òptics, tant com per *train* com per a codificació, s'ha seguit el treball realitzat anteriorment [3 - 5] i les notes han sigut anotades respecte de la seua posició *geomètrica* dins del tetragrama.

A la codificació geomètrica s'han emprat, bàsicament, dos símbols en minúscula per representar les notes: `l` i `s`. Indiquen si una nota ha sigut escrita en una línia o en un espai, respectivament. A més, s'afig un número a la dreta de cada símbol per especificar la seua posició relativa al tetragrama. La figura 4.2 mostra un exemple d'aquesta notació per a un fragment de tetragrama concret. A aquest fragment no s'han inclòs les codificacions de les notes que impliquen línies *addicionals* al tetragrama; això són símbols com a `l-1` o `l0` que es veuen de forma molt aïllada al llarg de tot corpus.

Dins de les partitures es poden trobar altres símbols rellevants musicalment: les claus i les alteracions (bemolls). Aquests també són representats a través de la codificació geomètrica.

Una clau es representa amb altre símbol en minúscula `c` o `f`, per a les claus de do i fa respectivament, seguits també d'un número que indica la línia en la qual han

sigut escrits. Per exemple: a la figura 4.2, `f3` representa la clau de fa en tercera; ha sigut escrita a la tercera línia.
El bemoll ♭ es representa directament com a `flat`, sense cap anotació numèrica.

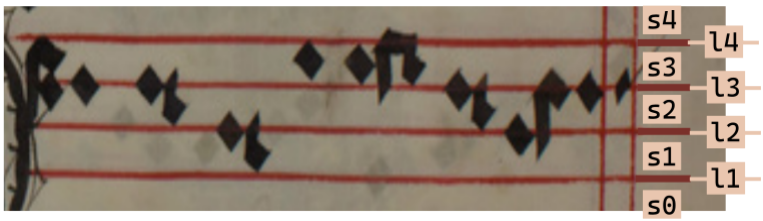


FIGURA 4.2: Fragment d'una imatge d'una de les pàgines del Vorau-253. Les posicions geomètriques verticals de les notes (quadrats negres, alguns amb traços ornamentals) són representades a partir dels símbols nou mostrats a la dreta de la imatge. La seqüència mostrada es representaria com a ⟨`f3,l3,l3,s2,l2,s1,s3,s3,l4,s3,l3,s2,l2,l3,l3`⟩

4.3 Particions experimentals

A continuació, es descriuen les particions del conjunt de dades Vorau-253 que s'han utilitzat per a l'entrenament (*train*), validació (*validation* o *val*) i prova (*test*) dels models i codificacions desenvolupats en aquest treball.

De forma general en cadascuna de les proves realitzades, la partició d'entrenament s'ha utilitzat per ajustar els paràmetres del model, mentre que la de validació ha estat emprada per a la selecció de models i l'ajustament dels hiperparàmetres. Finalment, la partició de prova s'ha reservat per a l'avaluació del model, oferint una estimació realista del seu comportament front dades no vistes.

Per a la realització dels diferents experiments, s'han creat arbitràriament tres grups bàsics de particions anomenats *Full*, *Half* i *Quarter* a partir d'ara. Cada grup difereix en la quantitat de dades assignades a les particions de *train* i *val*, mentre que la partició de *test* es manté constant en tots els casos, assegurant la comparabilitat dels resultats. A la taula 4.1 es poden veure les distribucions del nombre de pàgines i de línies per a cadascun dels grups de particions esmentats.

TAULA 4.1: Distribució del nombre de pàgines i línies del manuscrit en cadascun dels grups de particions per a l'experimentació realitzada.

	train		***val***		***test***	
	pàgines	línies	pàgines	línies	pàgines	línies
Full	250	2340	40	363		
Half	125	1133	20	197	200	1884
Quarter	60	542	20	197		

La variació en les particions d'entrenament i validació ha permès investigar la robustesa dels models davant diferents volums de dades, així com l'eficàcia de les codificacions BPE (Byte Pair Encoding) aplicades al Vorau-253, com s'explicarà en

detall en capítols posteriors d'aquesta memòria. Per tal d'aconseguir una distribució de les dades el més fidel possible a l'original, s'ha assegurat que en cada partició s'incloga al menys 1 vegada cadascun dels possibles símbols musicals del manuscrit. A la taula 4.2 es troba la distribució de símbols en cadascun dels grups de particions. Com es pot observar, tots els símbols han sigut vistos en *train*, encara que en *test* no s'hi troben.

TAULA 4.2: Distribució del nombre símbols en cadascun dels grups de particions per a l'experimentació realitzada.

	train			*test*
Symbol	***Full***	***Half***	***Quarter***	**totes**
c1	2	1	1	2
c2	125	54	41	105
c3	589	270	153	471
c4	1155	598	323	785
c5	128	113	49	2
f1	1	1	1	0
f2	2	2	1	0
f3	495	182	105	373
f4	36	24	13	10
flat	348	178	105	206
l0	14	14	14	31
l-1	1	1	1	1
l1	3096	1742	1000	2014
l2	10219	5003	2790	7403
l3	12529	6121	3364	8512
l4	4947	2785	1474	2479
l5	351	269	113	51
l6	3	2	2	0
s0	787	482	271	497
s-1	1	1	1	2
s1	5573	2827	1517	4012
s2	10049	5008	2821	6917
s3	6226	3297	1864	3669
s4	1290	821	444	453
s5	69	51	8	7

Capítol 5

Experimentació

Dins d'aquest capítol s'aprofundix en el treball d'experimentació realitzat. A través d'una sèrie d'experiments preliminars, es pretenen acotar els valors dels hiperparàmetres del *byte pair encoding* per a garantir una selecció eficient a utilitzar durant l'entrenament dels models òptics proposats.

Posteriorment, es descriu l'experimentació realitzada amb els models òptics, incloent les configuracions utilitzades i els entrenaments duts a terme. L'objectiu final és establir un marc sòlid per a les avaluacions realitzades al següent capítol.

5.1 Experimentació preliminar amb BPE

Abans de realitzar els diferents entrenaments de models òptics proposats, s'ha optat per fer una experimentació prèvia per acotar el rang de valors dels hiperparàmetres MIN_OCC i MAX_LEN de BPE que es provaran, ja que entrenar un model òptic per cada possible combinació d'aquestos resultaria prohibitiu temporalment, mentre que realitzar un anàlisi com el següent (basat només en el text i la forma de codificació de les dades) és factible dins de l'abast d'aquest treball.

D'aquesta forma, amb totes les dades disponibles d'entrenament (les 250 pàgines esmentades al capítol 4), s'han après diverses codificacions BPE i s'han avaluat les mètriques següents para cadascuna d'elles:

- **Grandària del vocabulari**: mesura la quantitat de *tokens* que formen el vocabulari après a partir de les dades d'entrenament.
- **IoU dels vocabularis** (*intersection over union*) o índex Jaccard[1] s'ha extret de la següent forma: a partir del vocabulari après amb les dades de *train*, s'han codificat les dades de *test*. Llavors, s'ha extret el vocabulari de *tokens* de les dades de *test* codificades i s'ha calculat l'índex de Jaccard entre ambdós vocabularis. D'aquesta forma, es mesura la similitud entre ells.

A les figures 5.1 i 5.2 es mostra la variació de les mètriques anteriors en les diferents codificacions BPE entrenades dins d'aquesta experimentació preliminar. Com es pot observar, en la majoria dels, a partir d'un MAX_LEN > 4 es comencen a estabilitzar les dues mètriques. Això indica que, quan es permeten *tokens* més llargs cada vegada, el vocabulari d'entrenament s'estabilitza. Això ocorre perquè no existeixen *tokens* tan llargs (vore taula 5.2) que apareguen 15 vegades o més.

[1] El càlcul d'aquest valor, per a dos vocabularis A i B és $J = \frac{|A \cap B|}{|A \cup B|}$

Per intentar tindre una representació de diversos valors de les mètriques escollides per a les codificacions dins dels entrenaments d'HMR experimentals que es realitzaran, i com que no resulta factible temporalment dur a terme un escaneig exhaustiu per als valors dels hiperparàmetres, s'ha decidit provar els entrenaments amb les codificacions BPE mostrades a la taula 5.1. Segons es pot vore més avall, a la secció 5.1.1, amb aquestes codificacions s'han intentat cobrir distints tamanys de vocabularis d'entrenament al mateix temps que l'índex IoU entre vocabularis es mantenia alt, ja que s'entén que les codificacions d'entrenament i prova seran més compatibles d'aquesta manera.

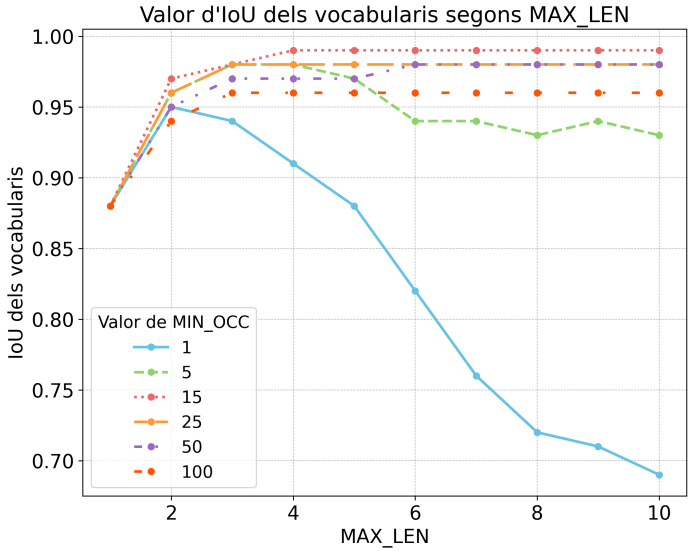


FIGURA 5.1: Variació de l'IoU dels vocabularis de *train* i *test* per a diferents configuracions BPE.

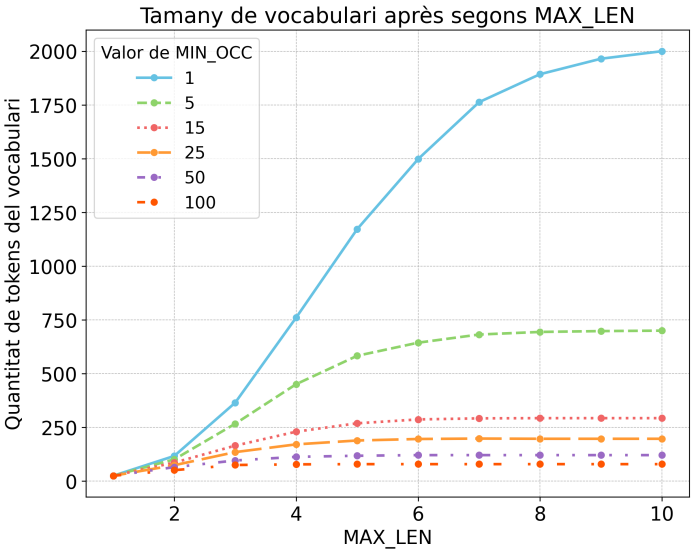


FIGURA 5.2: Variació de la grandària del vocabulari après amb diferents configuracions BPE.

TAULA 5.1: Hiperparàmetres de les codificacions BPE (*byte pair encoding*) escollides per experimentar el seu efecte en els entrenaments dels models òptics d'aquest treball.

Nom	MIN_OCC	MAX_LEN
Cod. 1	15	2
Cod. 2	15	4
Cod. 3	45	4
Cod. 4	50	10
Cod. 5	100	2

5.1.1 Detalls de les dades d'entrenament codificades

A partir dels hiperparàmetres BPE seleccionats, s'han codificat les dades d'entrenament tal i com s'ha detallat a la secció 3.2.2. Dins de la taula 5.2 es pot comprovar la distribució del vocabulari emprat en cada entrenament. Per entrenar els vocabularis BPE, s'ha emprat la partició destinada a l'entrenament del model en cada cas. És a dir que per codificar la partició d'entrenament *Half*, per exemple, s'han entrenat els diferents vocabularis BPE amb la mateixa partició (només d'entrenament) *Half*.

TAULA 5.2: Distribució de les dades d'entrenament per a cada model proposat, amb les configuracions BPE esmentades en 5.1. A la taula es mostren, a més dels hiperparàmetres BPE i les particions d'entrenament: la talla del vocabulari BPE, la longitud mitjana dels *tokens*, la mitja d'ocurrències per *token* i el total de *tokens* (no únics) d'entrenament.

Partició	MIN_OCC	MAX_LEN	Talla vocabulari	Longitud tokens (mitja)	Ocurrències tokens (mitja)	Total tokens
Full	15	2	86	1.7	406.9	34992
	15	4	230	2.9	102.8	23646
	45	4	122	2.6	215.6	26304
	50	10	121	2.7	216.0	26138
	100	2	51	1.5	715.4	36483
Half	15	2	76	1.7	237.4	18045
	15	4	160	2.71	80.4	12859
	45	4	83	2.2	177.7	1475
	50	10	79	2.3	188.8	14911
	100	2	46	1.5	416.7	19168
Quarter	15	2	66	1.6	152.2	10048
	15	4	115	2.5	65.8	7567
	45	4	67	2.1	130.0	8710
	50	10	65	2.1	135.1	8778
	100	2	39	1.4	282.4	11012

5.2 Entrenament dels models òptics

A la secció 2.2 al sistema d'HMR que es detalla s'aplica un model de llenguatge després d'entrenar el model òptic per millorar la qualitat del reconeixement, malgrat això, en el cas d'aquest treball fi de màster, s'ha preferit no incorporar-lo en l'etapa d'experimentació. La decisió de centrar-se exclusivament en l'avaluació del model òptic es fonamenta en la voluntat d'aprofundir en l'anàlisi i comprensió del comportament dels entrenaments dels models òptics quan s'empra la codificació BPE.
La tasca d'estudi de l'aplicació dels models de llenguatge en aquestos casos es planteja (es vorà al capítol 7) com a desenvolupament futur.

5.2.1 Configuracions de l'experimentació

A aquest treball, tots els models s'han entrenat amb una unitat NVIDIA GeForce RTX 2080. En el cas dels models amb BPE, s'ha seguit un entrenament amb *tokens* BPE dispersos (vore secció 3.2.3) en tots els casos. Amb la *Cod. 5* de la taula 5.1 també s'ha experimentat amb *tokens* BPE unificats, com es vorà més endavant al capítol 6 d'avaluació.

Els paràmetres d'entrenament, tant per als casos amb codificació BPE com per als d'entrenament *símbol a símbol* (experiments base), han sigut els següents:

- **Optimitzador:** Adam amb factor d'aprenentatge LR (*Learning Rate*) de 0.0003.
- ***Learning rate scheduler*:** (planificador del LR) basat en la pèrdua de validació, amb un factor del 0.1 i una paciència de 20 èpoques (*epochs*).
- ***Early stopping*:** (aturada anticipada) amb paciència de 40 èpoques.
- **Grandària de *batch*:** 8 mostres.

Capítol 6

Avaluació i resultats

Una vegada s'han realitzat tots els entrenaments, amb la finalitat de mesurar la qualitat dels sistemes proposats, és necessari avaluar-los.

En primer lloc, per a tots els models per igual, s'ha decodificat (vore secció 2.2.3) el conjunt de *test* descrit al capítol 4. Per tal de fer comparables els resultats dels models entrenats amb codificació BPE amb els models base (entrenats *símbol a símbol*), que anomenarem ORG (originals), s'han hagut de codificar les transcripcions produïdes amb aquestos models ORG per a cada tipus de codificació BPE proposada (apresa amb les dades d'entrenament corresponents).

6.1 Mètriques d'avaluació

Per mesurar el la qualitat dels sistemes, s'han emprat tres mètriques conegudes i comunes en l'àmbit de l'HTR [37]: taxa d'error a nivell de caràcter (CER[1]), taxa d'error a nivell de paraula (WER[2]) i taxa d'error a nivell de frase, línia, o tetragrama en aquest cas (SER[3]).

La WER d'una seqüència de símbols y (una línia) respecte d'un altra de referència x es defineix com a la distància d'edició entre x i y, normalitzada per la longitud de la seqüència de referència $n = |x|$:

$$\mathrm{WER}(x, y) = \frac{d(x, y)}{n} \equiv \frac{i + s + d}{c + s + d} \tag{6.1}$$

a on i, s, d són el mínim nombre d'insercions, substitucions i eliminacions, respectivament, necessitades per transformar x en y i c és el nombre d'elements correctes (que no necessiten edicions).

La CER es defineix de forma similar però assumint que n és el nombre total de caràcters en x, en aquest cas símbols musicals aïllats, i i, s, d, c són operacions d'edició i elements correctes a nivell de caràcter també.

La SER és, simplement, el quocient entre la quantitat de línies musicals (tetragrames) que no coincideixen exactament amb la referència i el nombre total de línies.

[1] De l'anglès *Character Error Rate*
[2] De l'anglès *Word Error Rate*
[3] De l'anglès *Sentence Error Rate*

6.2 Resultats

Per decidir quin tipus de format de les dades d'entrenament seguir (amb *tokens* dispersos o unificats, vore secció 3.2.3), s'han entrenat sis models diferents. Un per tipus de format de les dades i partició d'entrenament. A la taula 6.1 es mostren els resultats d'aquestes avaluacions.

TAULA 6.1: Resultats preliminars de CER, WER i SER per a dos entrenaments de models òptics: un amb *tokens* dispersos (forma `<space> s3 s2 s1 <space>`) i altre amb *tokens* unificats (forma `<space> s3#s2#s1 <space>`) amb diferents particions d'entrenament i els hiperparàmetres de la codificació BPE de MIN_OCC = 100 i MAX_LEN = 2. L'interval de confiança al 95% per al CER arriba fins a ± 0.22, el del WER a ± 0.35 i el del SER a ± 2.26.

Partició	Entrenament	CER (%)	WER (%)	SER (%)
Full	dispers	**1.9**	**4.8**	**23**
	unificat	2.1	5.0	27
Half	dispers	**3.5**	**7.7**	**40**
	unificat	4.5	8.2	50
Quarter	dispers	**4.3**	**8.7**	**50**
	unificat	5.0	8.3	53

Com que per a tots els casos d'entrenaments de la taula 6.1, les xifres de CER, WER i SER són menors (millors resultats) quan s'ha seguit el format de *tokens* dispersos, és aquest tipus de format el que s'ha emprat a la resta d'entrenaments que s'avaluen a continuació.

Finalment, a la taula 6.2 es mostren els resultats finals de CER, WER i SER en *test* dels models entrenats. S'observa que hi ha un model per tipus partició d'entrenament i tipus de codificació BPE realitzada. A més, per cadascun d'aquestos models (BPE a la taula), hi ha resultats per al seu model base corresponent (ORG a la taula).

Les dades de prova amb les quals s'han avaluat els models, s'han codificat d'acord amb el BPE de *train* corresponent. D'aquesta forma, en cada bloc de cada grup de particions (fileres BPE i ORG consecutives) la referència de *test* és la mateixa.

TAULA 6.2: Resultats finals de CER, WER i SER dels entrenaments amb *tokens* dispersos per a diferents particions d'entrenament i codificacions BPE (distints valors dels hiperparàmetres). L'interval de confiança al 95% per al CER arriba fins a ± 0.32, el del WER fins a ± 0.74, menys en el casos de *Full* que arriba només fins a ± 0.52, i el del SER fins a ± 2.24.

Partició	MIN_OCC	MAX_LEN	Tipus	CER (%)	WER (%)	SER (%)
Full	15	2	BPE	3.2	20.1	33
			ORG	2.1	10.0	28
	15	4	BPE	**1.6**	11.5	**22**
			ORG	2.1	9.3	28
	45	4	BPE	**1.6**	7.9	**22**
			ORG	2.1	7.0	28
	50	10	BPE	**1.5**	**6.0**	**21**
			ORG	2.1	6.8	28
	100	2	BPE	**1.9**	**4.8**	**23**
			ORG	2.1	5.3	28
Half	15	2	BPE	11.1	37.7	59
			ORG	3.4	12.8	40
	15	4	BPE	4.4	33.6	44
			ORG	3.4	12.5	40
	45	4	BPE	3.5	13.7	44
			ORG	3.4	9.3	40
	50	10	BPE	4.0	11.9	**40**
			ORG	3.4	9.1	40
	100	2	BPE	3.5	7.8	**40**
			ORG	3.4	7.0	40
Quarter	15	2	BPE	5.7	29.8	59
			ORG	4.9	15.1	56
	15	4	BPE	**4.0**	27.8	**45**
			ORG	4.9	17.4	56
	45	4	BPE	**3.5**	**13.1**	**44**
			ORG	4.9	13.7	56
	50	10	BPE	**3.6**	**12.0**	**43**
			ORG	4.9	13.3	56
	100	2	BPE	**4.3**	**8.7**	**50**
			ORG	4.9	9.4	56

6.3 Comentari

Una vegada s'han avaluat les diferents mètriques proposades, es poden extreure diverses conclusions sobre els resultats obtinguts en l'experimentació.

Com s'ha avançat anteriorment a aquesta memòria, s'observa que les taxes d'error tant a nivell de caràcter, com de paraula i línia empitjoren en tots els casos ORG a mesura que es decrementa la quantitat de dades d'entrenament. Açò no és d'estranyar ja que una menor quantitat de dades limita la capacitat del model per generalitzar correctament i capturar les variacions presents al manuscrit.

Per als models entrenats amb només 60 pàgines (*Quarter*), menys en el cas de la *Cod. 1* (MIN_OMM $=$ 15 MAX_LEN $=$ 2), es pot afirmar amb el 95% de confiança que els models milloren el seu reconeixement a nivell de caràcter i línia (CERs i SERs més baixos), així com també millora el seu reconeixement en tots els aspectes mesurats per a totes les codificacions, menys *Cod. 1* i *Cod. 2* (MIN_OMM $=$ 15 MAX_LEN $=$ 4). El mateix comportament es pot observar també per als models entrenats amb la partició *Full*.

A primera vista resulta anòmal el fet que amb algunes de les codificacions BPE (*Cod. 1*, *Cod. 2* i *Cod. 4*) s'aconseguisca millor CER amb la partició *Quarter* que amb *Half*, ja que la primera conté menys dades d'entrenament que la segona. Açò crida l'atenció especialment en la millora del CER de la *Cod. 1*, a on passa d'un 11.1% de CER amb *Half* a un 5.7% amb *Quarter*.

Com que la partició d'entrenament *Quarter* no està necessàriament inclosa a la *Half*, ja que les dades han sigut barrejades abans de dividir-les, es requeriria de més experimentació per confirmar l'anomalia i tractar de trobar les causes.

Els resultats mostren que la configuració de codificació amb paràmetres MIN_OCC $=$ 50 i $MAX\_LEN = 10$ presenta el millor rendiment general en la partició *Full*, aconseguint els millors valors de CER (1.5%), WER (6.0%) i SER (21%), encara que el WER no siga comparable amb la resta d'experiments amb BPE.

D'altra banda, els resultats de WER només es poden comparar per a cada parell ORG-BPE ja que en cada cas, al haver entrenat amb diferents particions els vocabularis BPE, el total de paraules sobre el qual s'extreu la mètrica varia.

Malgrat que en la majoria dels casos, el WER de BPE no supere al ORG, la qual cosa es pot entendre com que els models no han pogut arribar a aprendre els *tokens* de forma correcta, sí que es millora el reconeixement a nivell de caràcter. És especialment destacable en els casos de la partició *Quarter*. Açò podria interpretar-se com que, gràcies a la informació i el context que ofereixen les agrupacions de símbols musicals en *paraules musicals*, els models són capaços d'aprendre millor els símbols aïllats (com ocorre a HTR habitualment).

6.3.1 Implicacions dels Resultats

Els resultats obtinguts posen de manifest la importància de l'ús de codificacions adequades per a millorar el reconeixement òptic de partitures manuscrites. Aquests resultats suggereixen que la utilització de BPE com a mètode per a generar *paraules musicals* podria ser una via prometedora per a capturar millor la informació contextual dels símbols musicals. Malgrat que encara quede marge de millora, els resultats de taxes d'error obtinguts a nivell de caràcter indiquen que aquesta tècnica pot oferir una base sòlida per a futures investigacions.

A més, la presència d'anomalies, com la millora del CER en la partició *Quarter* en comparació amb *Half* en algunes codificacions, subratlla la necessitat de dur a terme una anàlisi més profunda d'aquest comportament. És possible que factors com la diversitat de les mostres seleccionades per a cada partició estiguen influenciant aquests resultats, i una millor comprensió d'aquestes variables podria portar a noves estratègies d'entrenament més robustes i efectives.

Aquestes troballes també suggereixen que caldria explorar altres estratègies per obtindre els vocabularis BPE, incloent la variació dels valors d'hiperparàmetres com MIN_OCC i MAX_LEN, així com l'anàlisi lingüístic dels vocabularis apresos, per tal d'identificar configuracions que maximitzen el rendiment dels models, també en situacions amb dades limitades.

En conjunt, l'anàlisi realitzada en aquest estudi serveix com a punt de partida per a futures investigacions orientades a la millora del reconeixement de partitures manuscrites, obrint noves vies per a la incorporació de models de llenguatge que puguen complementar els resultats òptics, com es concisa a continuació en la memòria.

6.4 Sobre les codificacions BPE

Per tractar d'entendre el comportament dels models, és convenient aprofundir en les dades que han sigut utilitzades per entrenar-los, això són les dades codificades amb BPE.

Quan s'entrena el BPE, s'obté un vocabulari amb el qual es codifiquen les dades. Amb l'objectiu de trobar *paraules musicals* que es puguen considerar elements bàsics d'un *llenguatge musical* i a partir dels vocabularis entrenats, una mesura interessant per tractar de caracteritzar aquest llenguatge és la seua corba Zipf.

La llei de Zipf és una propietat estadística de les llengües naturals. Aquesta estableix una relació entre les paraules més freqüents i la quantitat de vegades que apareixen en comparació amb la resta, indicant que la paraula més comuna es repeteix el doble de vegades que la segona, el triple que la tercera i així successivament.

Aquesta proporció inversa entre la freqüència d'ús d'una paraula i el rang ocupat en l'ordre d'aparicions de la paraula es pot observar en la majoria de llengües naturals estudiades [40]. D'aquesta manera, és possible, per a un vocabulari donat, treure la seua corba Zipf a partir de representar la freqüència d'aparició de cada *token* segons la seua posició en el rang de major a menor quantitat d'ocurrències. Aquest concepte ha sigut cas d'estudi en altres àmbits relacionats amb HTR [36] o inclús

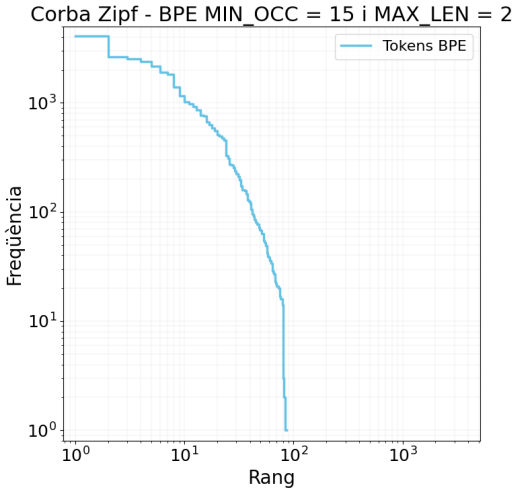
amb música [15, 21].

A partir de la base que amb cadascuna de les codificacions BPE proposades s'ha modelat un *llenguatge musical*, a la figura 6.1 s'han representat les corbes Zipf per a tots els vocabularis BPE entrenats amb la partició d'entrenament *Full*, que mostren el comportament del llenguatge corresponent.

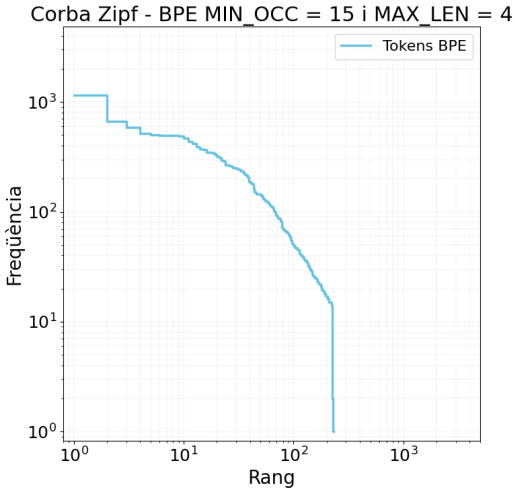
Les corbes Zipf representades no s'allunyen molt de les corbes de la majoria dels llenguatges naturals escrits o parlats [36]. Totes semblen tindre un tram prou clar amb pendent ≈ -1, que és un factor clau a la llei Zipf. De la resta de trams de les corbes, la causa de la caiguda abrupta al final és evident: el fet de forçar un límit inferior de les ocurrències dels *tokens* amb els entrenaments BPE (a través de l'hiperparàmetre MIN_OCC). D'aquesta forma quedaran pocs (o cap) *tokens* amb freqüències baixes, la qual cosa fa que la corba caiga en picat a partir del llindar de freqüència determinat.

Seria necessària més experimentació per poder observar una tendència entre l'adequació dels vocabularis BPE modelats amb els llenguatges naturals i la qualitat dels sistemes d'HMR entrenats amb aquestos vocabularis.

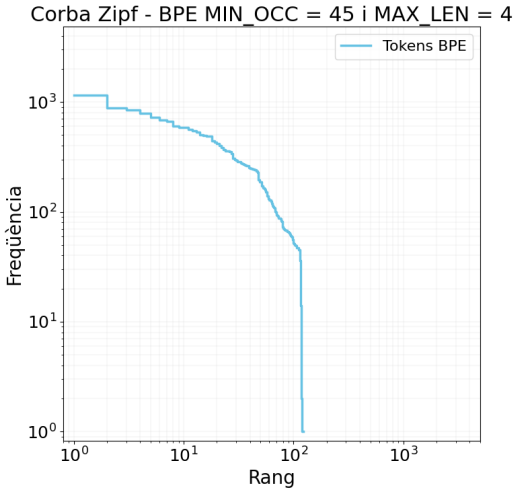
FIGURA 6.1: Corbes Zipf dels vocabularis de les cinc codificacions BPE emprades en l'experimentació per al conjunt de dades d'entrenament *Full*.



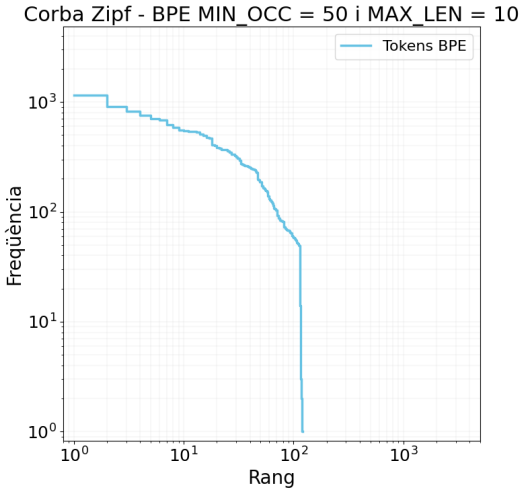
(A) Corba Zipf per al vocabulari de *Cod. 1*



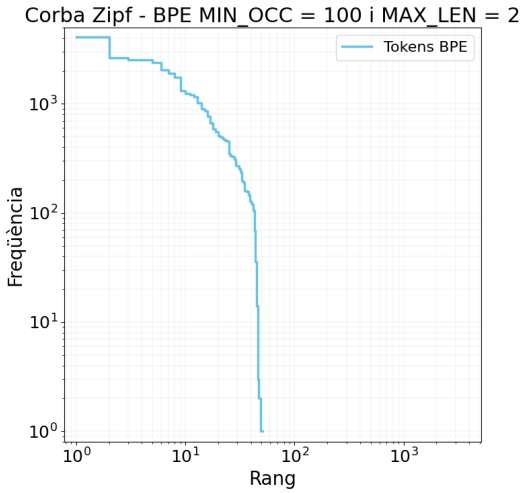
(B) Corba Zipf per al vocabulari de *Cod. 2*



(C) Corba Zipf per al vocabulari de *Cod. 3*



(D) Corba Zipf per al vocabulari de *Cod. 4*



(E) Corba Zipf per al vocabulari de *Cod. 5*

Capítol 7

Cloenda: perspectiva final

Al llarg d'aquest treball, s'ha analitzat l'impacte de l'ús del *byte pair encoding* (BPE) per al reconeixement de partitures manuscrites antigues a partir del cas d'estudi del manuscrit de Vorau-253 i s'han explorat diferents escenaris d'entrenament amb dades limitades. Els objectius establerts al començament de la investigació s'han assolit amb èxit, demostrant com la codificació BPE pot millorar el reconeixement de símbols musicals i oferint una base sòlida per a la futura investigació en l'àmbit de recuperació d'informació dins d'aquest tipus de manuscrits. A més, s'ha desenvolupat un marc experimental que facilita futurs estudis en la matèria, obrint noves perspectives d'investigació en el camp de la digitalització i preservació de documents històrics.

7.1 Limitacions i possibles millores

Tot i els objectius assolits, com a autocrítica, aquest treball presenta certes limitacions (degudes en part a causa del temps i els recursos disponibles) les quals ofereixen oportunitats per a millores. Entre aquestes limitacions es destaca:

- La manca d'una optimització detallada dels paràmetres d'entrenament de les xarxes segons el tipus de codificació emprada. Aquest aspecte podria haver permès una major eficiència en l'entrenament i millors resultats en la precisió de reconeixement.
- La limitació en l'exploració de codificacions BPE. Tot i haver-se provat diverses configuracions, una exploració més àmplia d'aquestes podria haver permès identificar configuracions més òptimes o més adaptades als diferents conjunts de dades.
- L'ús de particions limitades del conjunt de dades d'entrenament. Tot i que s'han obtingut conclusions significatives, una major variabilitat en les particions hauria permès observar tendències més clares en el comportament dels models segons la quantitat i el tipus de dades disponibles.

7.2 Conclusions extretes

Els resultats d'aquest estudi han confirmat que els objectius principals han estat assolits. El rendiment del sistema de reconeixement òptic de partitures ha mostrat una evolució positiva, en alguns casos, gràcies a la utilització del BPE, millorant el reconeixement de símbols musicals, també en situacions amb dades limitades. Així mateix, s'ha establert una base sòlida per a la investigació sobre recuperació d'informació musical dins dels manuscrits antics.

Tot i que els resultats mostren una millora clara en les mètriques de reconeixement, sobre tot a nivell de caràcter i línia, també s'han identificat anomalies que requereixen d'una anàlisi més profunda. Aquestes troballes han subratllat la necessitat de continuar investigant per a entendre millor la relació entre la codificació BPE i la qualitat del reconeixement òptic en diferents conjunts de dades.

7.3 Treball futur

Aquest treball obre noves vies d'estudi en diversos àmbits relacionats amb el reconeixement de partitures manuscrites. Una de les extensions més prometedores es tracta de la incorporació de models de llenguatge per complementar els resultats dels models òptics. Això podria portar a una millor comprensió del context musical i a una major precisió en el reconeixement.

Un aspecte relacionat amb açò que es pot plantejar també és com l'ús de diferents tipus de formats de les dades d'entrenament (en aquest treball només s'ha experimentat en detall amb les codificacions de *tokens* dispersos) afecta els resultats a l'hora d'aplicar els models de llenguatge (LM). Pot ser que les dades que ofereixen millors resultats òptics no siguin les més adequades per al LM, atès que unes es basen en *tokens* que són paraules dels vocabularis entrenats i altres en la predicció dels espais segons aquestes codificacions.

Altra línia futura de recerca podria incloure realitzar la *tokenització* dels vocabularis BPE de forma distinta, no només a partir dels *tokens* més llargs possibles d'esquerra a dreta, sinó també explorant altres formes de segmentació.

Seria interessant també desenvolupar estudis lingüístics detallats dels vocabularis BPE per adaptar-los a "comportaments" de les llengües naturals, amb l'objectiu d'explorar com açò afecta el rendiment dels models de reconeixement. Açò, no només per a BPE sinó per a altres tipus de codificacions possibles a més dels símbols musicals aïllats.

En conjunt, aquest treball estableix una base sòlida per a la investigació en el camp del reconeixement de partitures manuscrites, proporcionant eines per a la digitalització de documents històrics i noves vies per a la investigació futura.

Apèndix A

Objectius de Desenvolupament Sostenible (ODS)

A.1 Relació del treball amb els Objectius de Desenvolupament Sostenible (ODS) de l'agenda 2030

TAULA A.1: Grau de relació del treball amb els Objectius de Desenvolupament Sostenible (ODS).

Objectius de Desenvolupament Sostenible	**Alt**	**Mitjà**	***Baix***	***No procedeix***
ODS 1. Fi de la pobresa.				X
ODS 2. Fam zero.				X
ODS 3. Salut y benestar.				X
ODS 4. Educació de qualitat.			X	
ODS 5. Igualtat de gènere.				X
ODS 6. Aigua neta i sanejament.				X
ODS 7. Energia assequible i no contaminant.				X
ODS 8. Treball decent i creixement econòmic.			X	
ODS 9. Indústria, innovació i infraestructures.	X			
ODS 10. Reducció de les desigualtats.				X
ODS 11. Ciutats i comunitats sostenibles.				X
ODS 12. Producció i consum responsables.				X
ODS 13. Acció pel clima.				X
ODS 14. Vida submarina.				X
ODS 15. Vida d'ecosistemes terrestres.				X
ODS 16. Pau, justícia i institucions sòlides.				X
ODS 17. Aliances per aconseguir objectius.				X

Descripció de l'alineament del TFM amb els ODS amb un grau de relació més alt

Aquest treball de fi de màster es pot inscriure dins dels Objectius de Desenvolupament Sostenible (ODS), amb una relació particular amb l'objectiu número 9, que fa referència a la indústria, la innovació i les infraestructures. Així mateix, pot establir-se una connexió, encara que en menor grau, amb l'objectiu número 4 (educació de qualitat) i l'objectiu número 8 (treball digne i creixement econòmic).

D'una banda, aquest projecte ha permès delinear futurs camps de recerca i investigació que contribueixen al progrés tecnològic i científic. Aquestos aspectes no es limiten només a la tecnologia, sinó que també poden representar una aportació significativa al desenvolupament científic en el camp de la investigació històrica, ja que el treball realitzat aporta coneixements per al desenvolupament i millora d'una eina útil per a l'anàlisi de partitures musicals antigues.

Bibliografia

[1] T Bluche. *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur*. 2015.

[2] D Byrd i J G Simonsen. "Towards a standard testbed for optical music recognition: Definitions, met- rics, and page images". A: *Journal of New Music Research* 44.3 (2015), pàg. 169 - 195.

[3] Jorge Calvo-Zaragoza, Alejandro H. Toselli i Enrique Vidal. "Handwritten Music Recognition for Mensural notation with convolutional recurrent neural networks". A: *Pattern Recognition Letters* 128 (2019), pàg. 115 - 121. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2019.08.021. URL: https://www.sciencedirect.com/science/article/pii/S0167865519302338.

[4] Jorge Calvo-Zaragoza, Alejandro H Toselli i Enrique Vidal. "Probabilistic Music-Symbol Spotting in Handwritten Scores". A: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE. 2018, pàg. 558 - 563.

[5] Jorge Calvo-Zaragoza et al. "Music symbol sequence indexing in medieval plainchant manuscripts". A: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pàg. 882 - 887.

[6] Jifeng Dai et al. "R-FCN: Object detection via region-based fully convolutional networks". A: Cited by: 4766. 2016, 379 – 387. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85018938177&partnerID=40&md5=9317857555d943ae279ab7d47982b315.

[7] Mark Everingham et al. "The pascal visual object classes challenge: A retrospective". en. A: *Int. J. Comput. Vis.* 111.1 (2015), pàg. 98 - 136.

[8] Mark Everingham et al. "The pascal visual object classes challenge: A retrospective". en. A: *Int. J. Comput. Vis.* 111.1 (2015), pàg. 98 - 136.

[9] Nathan Fradet et al. "Byte pair encoding for symbolic music". A: *arXiv preprint arXiv:2301.11975* (2023).

[10] Philip Gage. "A new algorithm for data compression". A: *The C Users Journal* 12.2 (1994), pàg. 23 - 38.

[11] Xiaoxue Gao, Chitralekha Gupta i Haizhou Li. "Genre-Conditioned Acoustic Models for Automatic Lyrics Transcription of Polyphonic Music". A: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pàg. 791 - 795. DOI: 10.1109/ICASSP43922.2022.9747684.

[12] Alex Graves. "Long Short-Term Memory". A: *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pàg. 37 - 45. ISBN: 978-3-642-24797-2. DOI: 10.1007/978-3-642-24797-2_4. URL: https://doi.org/10.1007/978-3-642-24797-2_4.

[13] Alex Graves et al. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". A: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, 2006.

[14] Ximena Gutierrez-Vasques et al. "From characters to words: the turning point of BPE merges". A: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. de Paola Merlo, Jorg Tiedemann i Reut Tsarfaty. Online: Association for Computational Linguistics, abr. de 2021, pàg. 3454 - 3468. DOI: 10.18653/v1/2021.eacl-main.302. URL: https://aclanthology.org/2021.eacl-main.302.

[15] Martín Haro et al. "Zipf's law in short-time timbral codings of speech, music, and environmental sound signals". A: *Plos One* 7.3 (2012), e33993.

[16] Zhiqing Huang, Xiang Jia i Yifan Guo. "State-of-the-Art Model for Music Object Recognition with Deep Learning". A: *Applied Sciences* 9.13 (2019). ISSN: 2076-3417. DOI: 10.3390/app9132645. URL: https://www.mdpi.com/2076-3417/9/13/2645.

[17] Herman Kamper. "Word Segmentation on Discovered Phone Units With Dynamic Programming and Self-Supervised Scoring". A: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pàg. 684 - 694. DOI: 10.1109/TASLP.2022.3229264.

[18] Reinhard Kneser i Hermann Ney. "Improved backing-off for M-gram language modeling". A: vol. 1. Cited by: 1152. 1995, 181 – 184. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0028996876&partnerID=40&md5=664ab3b89cd5f8c3770d7288fc9aa64c.

[19] Adarsh Kumar i Pedro Sarmento. "From words to music: A study of subword tokenization techniques in symbolic music generation". A: *arXiv preprint arXiv:2304.08953* (2023).

[20] Paul Lascabettes i Isabelle Bloch. "Discovering Repeated Patterns From the Onsets in a Multidimensional Representation of Music". A: *International Conference on Discrete Geometry and Mathematical Morphology*. Springer. 2024, pàg. 192 - 203.

[21] Juan I Perotti i Orlando V Billoni. "On the emergence of Zipf's law in music". A: *Physica A: Statistical Mechanics and its Applications* 549 (2020), pàg. 124309.

[22] Daniel Povey et al. "The Kaldi speech recognition toolkit". A: *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. 2011.

[23] Laurent Pugin. "Optical music recognition of early typographic prints using Hidden Markov Models". A: Cited by: 57. 2006, 53 – 56. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-36348985923&partnerID=40&md5=905c6d1ff813a1c801deb5bffdbee3df.

[24] Lorenzo Quirós et al. *Vorau Abbey library Cod. 253 dataset for Document Layout Analysis*. 2021. DOI: 10.5281/ZENODO.5443257.

[25] Lawrence R Rabiner i Biing-Hwang Juang. *Fundamentals of Speech Recognition*. en. Philadelphia, PA: Prentice Hall, 1993.

[26] Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". A: vol. 2015-January. Cited by: 29973. 2015, 91 – 99. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84960980241&partnerID=40&md5=18aaa500235b11fb99e953f8b227f46d.

[27] George Retsinas et al. "Best Practices for a Handwritten Text Recognition System". A: *Document Analysis Systems*. Ed. de Seiichi Uchida, Elisa Barney i Véronique Eglin. Cham: Springer International Publishing, 2022, pàg. 247 - 259. ISBN: 978-3-031-06555-2.

[28] Christian Reul et al. "OCR4all—An open-source tool providing a (semi-) automatic OCR workflow for historical printings". A: *Applied Sciences* 9.22 (2019), pàg. 4853.

[29] Jaime dos Santos Cardoso et al. "Staff Detection with Stable Paths". A: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.6 (2009), pàg. 1134 - 1139. DOI: 10.1109/TPAMI.2009.34.

[30] Rico Sennrich, Barry Haddow i Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". A: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. de Katrin Erk i Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, ag. de 2016, pàg. 1715 - 1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

[31] Rico Sennrich, Barry Haddow i Alexandra Birch. "Neural machine translation of rare words with subword units". A: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.

[32] Elona Shatri i Gyorgy Fazekas. "Optical music recognition: State of the art and major challenges". A: *arXiv preprint arXiv:2006.07885* (2020).

[33] Baoguang Shi, Xiang Bai i Cong Yao. "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition". A: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.11 (2017). Cited by: 2039; All Open Access, Green Open Access, 2298 – 2304. DOI: 10.1109/TPAMI.2016.2646371. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85032274465&doi=10.1109%2fTPAMI.2016.2646371&partnerID=40&md5=d090b135e9ef6fc17e285131fb5b3ff0.

[34] Nitish Srivastava et al. "Dropout: A simple way to prevent neural networks from overfitting". A: *Journal of Machine Learning Research* 15 (2014). Cited by: 30960, 1929 – 1958. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84904163933&partnerID=40&md5=b865fd654b3befc5d829dbe5d42b80c3.

[35] Alejandro Héctor Toselli, Joan Puigcerver i Enrique Vidal. *Probabilistic indexing for information search and retrieval in large collections of handwritten text images*. Vol. 49. Springer Nature, 2024.

[36] Enrique Vidal i Alejandro H. Toselli. "Zipf Curves and Basic Text Analytics from Untranscribed Manuscript Images". A: *Document Analysis and Recognition - ICDAR 2024*. Ed. d'Elisa H. Barney Smith, Marcus Liwicki i Liangrui Peng. Cham: Springer Nature Switzerland, 2024, pàg. 271 - 288. ISBN: 978-3-031-70543-4.

[37] Enrique Vidal et al. "End-to-End page-Level assessment of handwritten text recognition". A: *Pattern Recognition* 142 (2023), pàg. 109695. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2023.109695. URL: https://www.sciencedirect.com/science/article/pii/S003132032300393X.

[38] Enrique Vidal et al. "Probabilistic finite-state machines - Part II". A: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.7 (2005). Cited by: 92; All Open Access, Green Open Access, 1026 – 1039. DOI: 10.1109/TPAMI.2005.148. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-22944443871&doi=10.1109%2fTPAMI.2005.148&partnerID=40&md5=9dd33112ea1d5ebb74cbefed80c5ee8f.

[39] Cuihong Wen et al. "A new optical music recognition system based on combined neural network". A: *Pattern Recognition Letters* 58 (2015), pàg. 1 - 7. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2015.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0167865515000392.

[40] Wikipedia contributors. *Llei de Zipf*. https://ca.wikipedia.org/w/index.php?title=Llei_de_Zipf&oldid=33880225. Accessed: 16-09-2024.

[41] Wikipedia contributors. *Optical music recognition*. https://en.wikipedia.org/w/index.php?title=Optical_music_recognition&oldid=1237718031. Accessed: 16-09-2024. Jul. de 2024.

[42] Philip F. Williams. A: *World Literature Today* 69.2 (1995), pàg. 433 - 434. ISSN: 01963570, 19458134. URL: http://www.jstor.org/stable/40151350 (cons. 10-09-2024).

[43] Matthew D. Zeiler i Rob Fergus. "Visualizing and Understanding Convolutional Networks". A: *Computer Vision - ECCV 2014*. Ed. de David Fleet et al. Cham: Springer International Publishing, 2014, pàg. 818 - 833. ISBN: 978-3-319-10590-1.