# Universal Grapheme Model for Medieval Latin Dictation

Firstname1 Surname1, Firstname2 Surname2, and Firstname3 Surname3

Affiliation1, Institute1, Address
www.website.org
{author1,author2}@institute1.org

**Abstract.** Medieval Latin pronunciation is considered to be uniform in the East-Central European (ECE) region. We present a medieval Latin charter dictation system which can be of great help for preserving language documents from the same era, since optical character recognition systems often fail to handle historic documents. In this paper, we introduce a medieval Latin dictation system for speakers from the ECE region to eliminate the efforts of digitizing medieval Latin charter data. In doing so, we train a Unified Grapheme Model (UGM) that can deal with cross-native-langauge variations. We show that our methods outperform our baseline system by ...

**Keywords:** dictation, latin, low resource speech recognition

## 1 Introduction

Apart from the two official pronunciations of Latin (classical and ecclesiastical), many regional pronunciations exist varying accross region and era. The third most known pronunciation group is the East-Central European (ECE) one, used for medieval Latin. Although the target pronunciation is considered to be uniform in this region, it is also has to be taken into account that the acoustic base of the different native languages varies, which can lead to different speakers pronouncing the same words differently. It also has to be noted, that apart from the variations in the pronunciations, orthographic and linguistic variations are also exhibited through regions. This raises the question of how to create a dictation system which has to deal with uniform pronunciations for speakers with different native languages reading linguistically different texts. We propose a system that is suitable for medieval Latin dictation for all speakers from the ECE region. The system we develop is a unified/joint system that can deal with both the variability in the speakers' pronunciations when speaking medieval Latin, and the grammatical/lexical variabilities of the texts. Our baseline system consists of separately trained grapheme (and phoneme) based acoustic models for the different languages in the ECE region. These separately trained models work good with their respective native speakers, but perform poorly with speakers of different native languages. We apply two different pronunciation modeling techniques to . The first one, dicussed in detail in Section 3.1, is based on the assumption that The second method we use is UGM (Unified Grapheme Modeling), where a joint/minimal/common grapheme inventory is established for all the languages paricipating in the joint acoustic model training. We describe this method in Section 3.2.

### 1.1 Related work

Similar work has been done for multi-dialectal languages such as Arabic in [1] where jointly trained acoustic models were outperformed by methods that unify dialect specific-acoustic models using knowledge distillation and multitask learning.

## 2 Data

### 2.1 Textual data

Textual data are scarce for medieval Latin. Additionally, most of the available sources mix local languages and Latin with no metadata to separate them. For the scope of this paper, we collected monolingual texts only.

**Alternate spellings** One interesting feature of the corpora is that they contain a significant number of spelling variants. Having alternate spellings with identical pronunciation introduces noise, and thus has a negative effect on recognition results. We obtained a unified spelling for these variants by favouring the more frequent variant in the corpus (e.g. *maiestati* to *majestati*). To detect the spelling variants we took all pairs in the pronunciation dictionary whose pronunciation were identical, and used context and expert knowledge to decide whether the pair of equivalent pronunciations are spelling variants or homophones. Where the decision was that they are spelling variants, the less frequent one was replaced by the more frequent one. Resolving spelling variants resulted in a more consistent corpus in terms of perplexity, and thus we expected better results with a model built from the corpus where spelling variants had been unified.

**Language model**

### 2.2 Speech data

## 3 Acoustic modeling

### 3.1 Grapheme to phoneme mapping

### 3.2 Unified grapheme modeling

## 4 Experimental results

### 4.1 Conclusions

## References

1. Waters, A., Bastani, M., Elfeky, M.G., Moreno, P., Velez, X.: Towards acoustic model unification across dialects. In: 2016 IEEE Workshop on Spoken Language Technology (2016)