Cross-Native-Language Medieval Latin Dictation

Firstname1 Surname1, Firstname2 Surname2, and Firstname3 Surname3

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

Abstract. 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. Our target era and geographical regions are medieval Latin, and documents originating from the Visegrad region, for speakers with Czech, Hungarian and Polish as their native languages. Our baseline systems are separately trained grapheme-based acoustic models for all the above three languages. We introduce two pronunciation modeling techniques to outperform the separately trained models. The first one is using grapheme-to-phoneme (G2P) mapping with Latin-specific pronunciation rules applied. The second one is training a Unified Simplified Grapheme (UGS) acoustic model that can deal with cross-native-langauge variations. We show that our methods outperform our baseline system, reducing the WER by ...% and ...% respectively.

Keywords: pronunciation modeling, 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 develop a models that are superior to the baseline. The first one, dicussed in detail in Section 3.2, is based on the assumption that The second method we use is USGM (Unified Simplified 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 3.3.

1.1 Related work

Similar work has been done for multi-dialectal languages such as Arabic in [5] 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

As part of out inquiry was to cover linguistic variability accross the ECE region, aquiring textual data posed a few challenges. First of all, textual data are scarce for medieval Latin, and texts originating from the ECE geographical region are even more scarce. 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.

Training data A smaller amount of in-domain data (medieval charters) were collected from [3] (Monasterium), with and overall of 480k tokens. These documents are originating from the Hungarian Kingdom, from 1000 to 1524 AD. To increase the vocabulary size of the language model, we collected a relatively larger (but still small, compared to state-of-the-art language models used in speech recognition) 1.3 token corpus from [4] (LatinLibrary). This corpus consists of literary and historical texts from the post-classical era. In spite of our efforts, at the time of writing this paper, we could not gather textual data from the age and area of the Kingdoms of Bohemia and Poland.

Test data Using independent sources three-three charters were selected from the Kingdoms of Bohemia (CZ), Hungary (HU) and Poland (PL), from around 1200-1300 AD, as development and test data. The dev set was used for evaluating the language model, and the test set to test the performance of our recognizers, by having them read out loud by historians fluent in medieval Latin.

Alternate spellings One interesting feature of the acquired corpora is that they contain a significant number of spelling variants. Having spelling variants in the corpus 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 (reducing it from 775 to 672), and reduced the OOV rate by 0.8%.

Language model The language models we built from the two corpora were estimated with the SRI Language Modeling toolkit (SRILM) [2] using modified Kneser-Ney smoothing method. After estimating the mixture parameter, linear interpolation was used.

The perplexity measures on the dev data showed that the Monasterium corpus originating from the time and era of the Hungarian Kingdom

2.2 Speech data

3 Acoustic modeling

For all the different pronunciation modeling methods, the acoustic models were trained as follows. Mel-Frequency Cepstrum $+$ Energy features were used with Linear Discriminant Analysis (LDA) + Maximum Likelihood Linear Transformation (MLLT), with a splice context of ± 4 frames, 10 ms of frame shift. 9×40 dimensional spliced up feature vectors served as input to the feedforward, 6 hiddenlayer neural network with pnorm [16] activation function. Prior to DNN training, a Gauss Mixture Model (GMM) pre-training was performed. Clustering and Regression Tree (CART) [1] was applied to obtain acrossword context dependent shared state phone (or graph) models and their time alignment. The number of senones (and so the size of the DNN softmax output layer) was between 7.000 and 11.000 depending on the nature of the training data. The size of the hidden layers was kept constantly on 2.000. A minibatch size of 512, an initial learning rate of 0.1, and final learning rate of 0.01 was applied in 20 epochs using the KALDI toolkit [1].

3.1 Grapheme

For our three separately trained baseline systems grapheme-based pronunciation models were used. It was based on the same principles described in detail in Section 3.3, namely mapping the graphemes not present in the Latin grapheme set to their normalized counterparts.

3.2 Grapheme to phoneme mapping (G2P)

The Czech and Hungarian phoneme-based acoustic models we used were trained with G2P mapping between orthographic transcriptions and native phonemes. The G2P rules were composed with the language model within a WFST framework [**?**]. But since we were using these models for recognizing medieval Latin speech spoken by native speakers of Czech, Hungarian and Polish, Latin-specific pronunciation rules also had to be implemented. These include a few context independent digraph mappings, and a few context dependent rewrite rules, summarized in Table 3.2 and Table 3.2 respectively, for both Czech and Hungarian.

Table 1. Latin digraph context-insensitive rewrite rules.

Digraph	ae	oe	ph	qu
CZ	e	oe	f	kv
HU	e	ø	f	kv

Table 2. Latin context-sensitive rewrite rules.

GR	c	c	ch	ch	gu	gu	ti	ti
PH	ts	k	h	k	gv	gu	tsi	ti
rule	cVP	cVNP	VC*ch	¬C*ch	guV	guC	tiV	tiC

3.3 Unified Simplified Grapheme Modeling

The second method we propose for cross-native-langauge Latin dictation is Unified Simplified Grapheme (USG) pronunciation modeling technique, which comes in play when joint acoustic models are being trained to support recognition across multiple languages.

Unified The joint acoustic model requires a unified grapheme inventory for the trainig, so that only thosegraphemes are in the model that are in the intersection of the different grapheme inventory sets of the training languages. Those letters that are not in this intersection are mapped to their normalized forms, e.g. it had a diacritic mark (acute, caron, etc.) on it, we mapped it back to its normalized form (`ř` to `r`, etc.).

Simplified Since the target dictation language was medieval Latin, the remaining unified grapheme set also had to be simplified to the Latin grapheme set, e.g. `ó` to `o`. Further than that, those graphemes that are non-native to Latin, and can straightforwardly mapped to a native Latin grapheme(s), were also replaced. These are mappings from `x` to `ks`, `y` to `i` and `w` to `v`. As a result, a unified and simplified grapheme inventory set was produced, compatible with medieval Latin. The USG units were then used as acoustic model units in the training.

4 Experimental results

Error analysis

4.1 Conclusions

In this paper, we presented two pronunciation modeling techniques for a cross-native-language medieval Latin dictation system to eliminate the efforts of digitizing medieval Latin charter data. With the objective of outperforming the separately trained grapheme-based models, we presented two approaches: an expert G2P modeling, and UGS modeling. The results showed...

Future research directions include acquiring a considerable amount of medieval speech and textual data.

Table 3. Polish grapheme-based baseline model. Size of acoustic model: 31 hours.

Speaker	CZ	HU	PL	Avr.
CZ	45.8	38.5	69.1	51.1
HU	74.3	48.2	68.8	63.8
PL	64.5	50.9	81.7	65.7
Avr.	61.5	45.9	73.2	60.2

Table 4. Czech Latin-specific G2P model. Acoustic model size: 76 hours.

Speaker	CZ	HU	PL	Avr.
CZ	39.3	28.8	47.4	38.5
HU	52	40	58.7	50.2
PL	94.1	67.3	97.2	86.2
Avr.	61.8	45.4	67.8	58.3

Table 5. Hungarian Latin-specific G2P model. Acoustic model size: 567 hours.

Speaker	CZ	HU	PL	Avr.
CZ	19.7	7.3	26.3	17.8
HU	25	25.4	20.2	23.5
PL	47.4	24.6	60.5	44.2
Avr.	30.7	19.1	35.7	28.5

Table 6. USG model of Czech (76 hours), Hungarian (112 hours) and Polish (31 hours).

Speaker	CZ	HU	PL	Avr.
CZ	24.8	17.9	36.1	26.3
HU	32.9	20.9	30.3	28
PL	57.2	38.2	78	57.8
Avr.	38.3	25.7	48.1	37.4

Table 7. USG model of Hungarian (112 hours), Polish (31 hours) and Roman (35 hours).

Speaker	CZ	HU	PL	Avr.
CZ	23.2	14.2	32.7	23.4
HU	23	14.6	27.5	21.7
PL	55.3	30	68.8	51.4
Avr.	33.8	19.6	43	32.2

Table 8. USG model of Czech (76 hours), Polish (31 hours) and Roman (35 hours).

Speaker	CZ	HU	PL	Avr.
CZ	20.8	13	35.2	23
HU	36.8	23.6	39.5	33.3
PL	55.9	32.7	67.9	52.2
Avr.	37.8	23.1	47.5	36.2

Table 9. USG model of Czech (76 hours), Hungarian (112 hours), Polish (31 hours) and roman (35 hours).

Speaker	CZ	HU	PL	Avr.
CZ	18.4	12.1	28.7	19.7
HU	20.4	14.6	25.7	20.2
PL	54.6	25.4	64.2	48.1
Avr.	31.1	17.4	39.5	29.3

References

1. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
2. Stolcke, A.: Srilm – an extensible language modeling toolkit. In: IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP) 2002. pp. 901–904 (2002)
3. `http://monasterium.net/mom/HU-PBFL/archive`: Monasterium.net archive
4. `http://www.thelatinlibrary.com/medieval.html`: Latin library archive
5. 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)