Language-Independent Acoustic Modeling for Medieval Latin

Firstname1 Surname1, Firstname2 Surname2, and Firstname3 Surname3

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

Abstract. A large vocabulary continuous speech recognition (LVCSR) system designed for dictation of medieval Latin language documents is introduced. Such language technology tool can be of great help for preserving Latin language charters from this era, as optical character recognition systems are often challenged by these historic documents. As corresponding historical research focuses on the Visegrad region, our primary aim is to make medieval Latin dictation available for texts and speakers of this region, concentrating on Czech, Hungarian and Polish. The baseline acoustic models we start with are monolingual grapheme-based ones - according to the language resources available - for these three languages. In one hand, the application of medieval Latin knowledge-based grapheme-to-phoneme (G2P) mapping from the source language to the target language resulted in significant improvement, reducing the WER by 17.3%. On the other hand, applying a unified-simplified grapheme (USG) inventory set for the three-language acoustic data set complemented with Romanian speech data yielded in competitive results - without using any target or source language G2P rules.

Keywords: G2P, medieval Latin, under-resourced speech recognition, unified simplified grapheme modeling

1 Introduction

The pronunciation of Latin texts mainly depends on the era and region of their origin [3]. Apart from the two widely studied classical and ecclesiastical pronunciation styles [1], regional pronunciations emerged after the classical era. One of these pronunciation groups is the east-central european [3] one, which uses roughly the same pronunciation rules, described in detail in Section 3.2. Although the target pronunciation is considered to be uniform for this group, it is also has to be taken into account, that the acoustic base of the different source 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 of Latin are also exhibited through regions.

This raises the question of how to create a speech recognition system which has to deal with pronunciation variations for native speakers of different languages reading linguistically different texts. We propose a system that aimes at the recognition of medieval Latin speech spoken by speakers from the Visegrad region. Therefore, it is

important to collect in-domain textual/language data for the language model from the relevant geographican regions and time. We describe the data acquisition process in section 2.1.

Our baseline system consists of separately trained grapheme-based acoustic models for the three Visegrad languages (Czech, Hungarian, Polish and Romanian. We apply two different acoustic/pronunciation modeling techniques to develop models that are superior to the baseline. The first one, dicussed in detail in Section 3.2, is a knowledge-based pronunciation modeling technique, where the source language phonemes are mapped to the target language phonemes. The second method we use is USG (Unified Simplified Grapheme) modeling, where a joint grapheme inventory is established for all the languages paricipating in the joint acoustic model training. We describe the USG method in Section 3.3. Evaluation of the baseline systems and both above approaches is presented in section 4.

1.1 Related work

Different adaptaion techniques have been proposed in [5] to train acoustic models from multiple source languages for a single target language where no training data was available. [2] gives a great overview on designing speech recognition systems for under-resourced languages. Similar work has been done for multi-dialectal languages such as Arabic in [12] where jointly trained acoustic models were outperformed by methods that unify dialect specific-acoustic models using knowledge distillation and multitask learning.

To our knowledge, no previous work has been done on medieval Latin speech recognition, nor on classical Latin for that matter.

2 Data

2.1 Textual data

As part of out inquiry was to cover linguistic variability accross the Visegrad region, aquiring textual data posed a few challenges. First of all, textual data are scarce for medieval Latin, and texts originating from this 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 [10] (Monasterium), with an 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 [11] (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, for development and test data. The dev set was used for evaluating the language model, and to test the performance of our recognizers. Both dev and test sets were 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. 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. We obtained a unified spelling for these variants by favouring the more frequent variant in the corpus (e.g. *maiestati* to *majestati*). 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) [6] 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 was indeed best fitting with the Hungarian subset of the test data with a perplexity of 82, and an OOV rate of 0.9%. Adding the LatinLibrary corpus increased the perplexity significantly, but reduced the OOV rate by 7% on the overall test data, as well as the WER, so we decided to use the interpolated language model.

2.2 Speech data

Training data For Czech and Hungarian the Speecon databases [9] and [8] and broadcast news speech data was used. For Polish, only broadcast news data [7] was available, comprising 31 hours of manually transcribed speech. The Romanian speech database we used for our expriments was collected for [7] consisting of 35 hours of broadcast news.

Test data Native speakers of Czech, Hungarian, Polish and Slovakian all of whom have experince with medieval Latin were asked to record the three dev and test sets described in Section 2.1. The recording conditions were accurately controlled: closetalking microphones, quiet, non reverberant acoustic environment, fluent, flawless speech, and at least 16 kHz, 16 bit (linear PCM) encoding. No instructions were given regarding the pronunciation, the speakers were using their expertise on medieval Latin pronunciation rules combined with their native language pronunciation. The overall length of the recorded test speech was around 30 minutes.

3 Acoustic modeling

Building an acoustic model for speech recognition requires long hours of trancribed speech. As of today (medieval) Latin is not spoken natively, and as to our knowledge, there is no recorded speech database. One obvious way to handle this problem is create a medieval Latin database; a proposition that requires lot of time, resources and trained speakers of medieval Latin. Another way of circumvent the lack of available speech data is to use speech data of spoken languages, preferably those ones whose native speakers are going to use the system.

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) [4] 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 [4].

3.1 Grapheme-based pronunciation modeling

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 Source-target grapheme to phoneme mapping (G2P)

As source language acoustic training data we used Czech and Hungarian phoneme-based acoustic models with G2P mapping from orthographic transcriptions to native phonemes. To develop a pronunciation dictionary for the target language medieval Latin, first we mapped source languages phonemes to target language phonemes using expert knowledge. As a second step, Latin-specific pronunciation rules also had to be mplemented. These include a few context independent digraph mappings, and a few context dependent rewrite rules, summarized in Table 1 and Table 2 respectively, for both Czech and Hungarian. Both languages fully cover the phoneme inventory of medieval Latin. The Latin alphabet we extracted from our corpora (see Section 2.1) consisted of 24 elements.

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. V: vowel, VP: palatal vowel, C: consonant, $*$: zero or any, ^: beginning of word, $[\hat{}\,stx]$: not s, t or x.

GR	c	c	ch	ch	gu	gu	ti	ti
PH	ts	k	h	k	gv	gu	tsi	ti
rule	cVP	VNP	VC*ch	^C*ch	guV	guC	$[\hat{}\,stx]$tiV	tiC

3.3 Unified Simplified Grapheme Modeling

The second method we used for improving language-independent speech recognition for medieval Latin was the Unified Simplified Grapheme (USG) pronunciation modeling technique. Our motivation with using this technique was two-fold:

1. Develop a target language acoustic model using different source languages.
2. Support recognition of medieval Latin spoken by speakers of a diverse native language background.

For that we used a joint acoustic model of Czech, Hungarian and Polish.

The joint acoustic model requires a unified grapheme inventory for the trainig, so that only those graphemes are in the model that are in the intersection of the different grapheme inventory sets of the training languages. Those graphemes 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.). Since the target 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, formally compatible with medieval Latin. The USG units were then used as acoustic model units in the training.

4 Experimental results

We conducted expeiments on medieval Latin, spoken by native speakers of three languages (Czech, Hungarian and Polish), where the test texts were originating from different regions, as described in Sections 3 and 2.1. The experiments were performed using two techniques, with the intention of improving recognition results on separately trained grapheme-based models, described in 3.1. Our experimental results showed that both proposed methods outperformed the baseline system.

It has to be noted that the experiments conducted on the Hungarian text test set yielded to the best reults with all models. This is due to the fact, that the in-domain part of the language model training data was originating from the Hungarian language region

see Section 2.1. On a related note, we also found that except for Polish, the baseline acoustic models were yielding the best results when testing with native speakers of the source language.

The best performing monolingual grapheme-base model results were that of Hungarian, with 45.8% overall WER (see Table 3) - this was the reference value when comparing the results.

Table 3. Word Error Rate (WER[%]) results for monolingual grapheme-based models of Czech (76 hours), Hungarian (112 hours) and Polish (31 hours).

	Speaker				
AM Language	CZ	HU	PL	SK	Avr.
CZ	53.6	74	94.1	45.7	66.8
HU	33.7	28.1	77.1	29.1	42
PL	65	68.3	73.1	51.1	64.4
RO	53.6	68	89.2	43.8	63.7

4.1 Source-target G2P mapping results

The results on the experiments with the knowledge-based pronunciation modeling technique, where the native phonemes of the source phoneme-based acoustic models were mapped to the target phonemes in the pronunciation dictionary, are in Table 4 for the source language Czech, and in Table 5 for the source language Hungarian. The Hungarian knowledge-based acoustic model results are by far the best ones (28.5% overall WER). It is worth mentioning that the Czech speakers achieve a surprisingly low 6.4% WER on the Hungarian text test set.

We did not include a Polish source phoneme-based acoustic model results in this paper as we were lacking an expert-based Polish G2P.

Table 4. WER of Latin-Czech source-target G2P model. Acoustic model size: 76 hours.

	Latin Test Text			
Speaker	CZ	HU	PL	Avr.
CZ	43.8	28.2	49.1	40.4
HU	52	40	58.7	50.2
PL	94.1	67.3	97.2	86.2
SK	30.3	30	44	34.8
Avr.	55	41.4	62.2	52.9

Table 5. WER of Latin-Hungarian source-target G2P model. Acoustic model size: 567 hours.

	Latin Test Text			
Speaker	CZ	HU	PL	Avr.
CZ	19.4	6.4	28	17.9
HU	25	25.4	20.2	23.5
PL	47.4	24.6	60.5	44.2
SK	20.4	9.1	22.9	17.5
Avr.	28	16.4	32.9	25.8

4.2 USG results

The results for the Czech-Hungarian-Polish joint acoustic model are in Table 6. We also tried adding a new source language, Romanian, to the joint USG acoustic model, which improved the results significantly, yielding in an overall 29.3% WER, with a best WER if 11.8% of the Czech speakers on the Hungarian test test, see in Table 7. The intution behind adding Romanian as a further source language is, that out of the four languages, it the most closely related one to Latin.

Additionaly, we also measured the WER on any combination of three of the four source langauges, to see how each source acoustic model contributes to the joint USG model. We found, that each language contributed almost evenly to the four-language USG model.

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

	Latin Test Text			
Speaker	CZ	HU	PL	Avr.
CZ	26	18.6	39.9	28.2
HU	32.9	20.9	30.3	28
PL	57.2	38.2	78	57.8
SK	22.4	16.4	28.4	22.4
Avr.	34.6	23.5	44.1	34.1

Table 7. WER of USG model of Czech (76 hours), Hungarian (112 hours), Polish (31 hours) and Romanian (35 hours).

	Latin Test Text			
Speaker	CZ	HU	PL	Avr.
CZ	20.4	11.8	30.7	21
HU	20.4	14.6	25.7	20.2
PL	54.6	25.4	64.2	48.1
SK	14.5	12.7	24.8	17.3
Avr.	27.5	16.1	36.4	26.7

The most striking results are the Czech speakers on the Hungarian text test set with both the knowledge-based and USG models. We had expected the Hungarian speakers to perform better with the Hungarian knowledge-based model and Hungarian text test set setting, but in fact the phoneme mapping masked the difference between mid-front /eː/ and open-front /ɛ/ in the pronunciation of the Hungarian speakers. In addion to that, they were pronouncing the named entities using their native pronunciation, which also increased the WER.

4.3 Conclusions

In this paper, we presented two pronunciation modeling techniques for a target-language independent medieval Latin speech recognizer to eliminate the efforts of digitizing medieval Latin charter data. Our goal was to build an acoustic model for medieval Latin, borrowing speech data from different source languages (Czech, Hungarian, Polish and ultimately Romanian). Our test set consisted of medieval Latin charters originating from different regions read by native speakers of the above languages. With the objective of outperforming the monolingual grapheme-based models, we presented two approaches: knowledge-based G2P modeling, and USG modeling. The results showed that both methods outperform by large the best baseline system. We concluded our work with comparing the results on a by-speaker-language and by-text-origin basis, where we found speaker-language differences affecting the results.

Future research directions include acquiring a considerable amount of medieval speech and textual data, as well as implementng a more refined G2P modeling using a global pheneme inventory set.

References

1. Allen, W.S.: Vox Latina: a guide to the pronunciation of classical Latin. Cambridge University Press Cambridge [Eng.] ; New York, 2d ed. edn. (1978)
2. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Communication 56, 85–100 (2014)
3. Encyclopedia, W.H.: Latin regional pronunciation (2007)
4. 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)
5. Schultz, T., Waibel, A.: Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Communication 31, 31–51 (2001)
6. Stolcke, A.: Srilm – an extensible language modeling toolkit. In: In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). pp. 901–904 (2002)
7. Tarjan, B., Mozsolics, T., Balog, A., Halmos, D., Fegyo, T., Mihajlik, P.: Broadcast news transcription in central-east european languages. In: 3rd IEEE International Conference on Cognitive Infocommunications. pp. 59–64 (2012)
8. `http://catalog.elra.info/product_info.php?products_id=1093`: Hungarian speecon database (2003)
9. `http://catalog.elra.info/product_info.php?products_id=1095`: Czech speecon database (2004)
10. `http://monasterium.net/mom/HU-PBFL/archive`: Monasterium.net archive
11. `http://www.thelatinlibrary.com/medieval.html`: Latin library archive
12. 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)