Unified Simplified Grapheme Acoustic Modeling for Medieval Latin LVCSR

Lili Szabó[1], Péter Mihajlik[2,3], András Balog[2], and Tibor Fegyó[1,3]

[1] SpeechTex
www.speechtex.com
{lili,tfegyo}@speechtex.com
[2] Thinktech
{mihajlik,abalog}@thinktech.hu
[3] Budapest University of Technology and Economics

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 materials. 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. On 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 Word Error Rate (WER) by 13.3%. On the other hand, applying a Unified Simplified Grapheme (USG) inventory set for the three-language acoustic data set complemented with Romanian speech data, resulted in a further 0.7% WER reduction - 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 also has to be taken into account, that the acoustic base of the different source languages varies, which can lead to various accents. It also has to be noted, that apart from the variations in the pronunciations, orthographic and grammatical 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 three of the Visegrad languages (Czech, Hungarian, Polish) complemented with the Romance language Romanian. We apply two different acoustic/pronunciation modeling techniques to develop models that are superior to the baseline. The first one, discussed 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 applied is a Unified Simplified Grapheme (USG) acoustic modeling approach, where a joint grapheme inventory is established for all the languages participating 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 adaptation techniques have been proposed in [5] to train acoustic models from multiple source languages for a single target language where training data was limited. [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. However, no approach is known for the authors where the graphemes of multiple languages are merged successfully and applied for acoustic modeling of a different language. 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 our inquiry was to cover linguistic variability across 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 difficult to obtain in electronic format. 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 (Latin) 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) 1.3M-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 a measurable amount of textual data from the age and area of the Kingdoms of Bohemia and Poland.

Test data Using independent sources, three charters were selected from the Kingdoms of Bohemia (CZ), Hungary (HU) and Poland (PL), from around 1200-1300 AD, as test data. The dev set was used for evaluating the language model, and to test the performance of the LVCSR approaches. The 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 word trigram 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 to merge the two language models.

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%. The perplexities measured on the Czech and Polish origin text sets were ranging from 500 to 3200. Adding the LatinLibrary corpus increased the perplexity significantly (up to 672), 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, the read part of Speecon database [9] was used, 76 hours in sum. For Hungarian, beyond Speecon [8], manually transcribed broadcast news (112 hours) and conversational speech data was used, altogether 567 hours. With the exception of the Hungarian knowledge-based model (described in Section 3.2), the 112-hour broadcast news set was used for training. For Polish, only broadcast news data [7] was available, comprising 31 hours of manually transcribed speech. The Romanian speech database used for the experiments was originally 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 - affected certainly by their native language. 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 transcribed 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 by creating a medieval Latin database; a proposition that requires lot of time, resources and trained speakers of medieval Latin. Another way of circumventing 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 feed-forward, 6 hidden-layer neural network with p-norm [4] activation function. Prior to DNN training, a Gaussian 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 acoustic models were used where pronunciation is modeled in an implicit way. The language-specific graphemes (e.g. ö, ń) that are not part of the Latin alphabet were trained, but not used in the recognition phase.

3.2 Source-target grapheme to phoneme mapping (G2P)

The source languages for the acoustic model training were Czech and Hungarian with G2P mapping from orthographic transcriptions to native phonemes. To develop a pronunciation dictionary for the target language medieval Latin, first we mapped source

language-phonemes to target language phonemes using expert knowledge. As a second step, Latin-specific pronunciation rules also had to be implemented. These include a set of context independent digraph mappings and 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, 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, ˆVP: everything but a palatal vowel, C: consonant, $*$: zero or any, ˆ: 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	cˆVP	VC*ch	ˆC*ch	guV	guC	$[\hat{}\,stx]$tiV	tiC

3.3 Unified Simplified Grapheme Acoustic Modeling

The second method used for improving speech recognition of medieval Latin - this time in a fully data driven way - was the Unified Simplified Grapheme (USG) acoustic modeling technique. Our motivation with using this technique was three-fold:

1. Develop a target language acoustic model using available language resources.
2. Support recognition of medieval Latin spoken by speakers of diverse native language background.
3. As the writing systems in the Visegrad region are originating from medieval Latin, we were aiming to validate the intuition that by unifying and simplifying the native graphemes, the deviations from the common ancestor cancel out.

We experimented with joint three- and four-language USG acoustic models of any combination of the four languages (Czech, Hungarian, Polish and Romanian). The joint acoustic model requires a unified grapheme inventory for the training. Our proposal was to simplify all special characters, i.e. those graphemes that had a diacritic mark (acute, caron, etc.) on them, were mapped back to their normalized form (e.g. `ř` to `r`, `í` to `i`, etc.). Table 3 contains examples for the unification/simplification process for all four languages. For the four languages an overall of 32 of such unifications/simplifications were made. 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 included 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.

Table 3. Simplification examples for the unified model.

Language	CZ	HU	PL	RO
Orthographic	řekl	őz	miś	apă
USG	rekl	oz	mis	apa

4 Experimental results

We conducted experiments on medieval Latin, spoken by native speakers of four languages (Czech, Hungarian, Polish and Slovakian), where the test texts were originating from different regions, as described in Sections 3 and 2.1. The best performing monolingual grapheme-based model results were that of Hungarian, with 34.6% overall WER (see in Table 4), possibly because of the larger training data - this was the reference value when comparing the results. On a related note, we also found that except for Czech, each monolingual grapheme-based acoustic model had the best performance over its own test set.

Table 4. Word Error Rate (WER[%]) results for monolingual grapheme-based acoustic models of Czech, Hungarian, Polish and Romanian (CZ, HU, PL, RO).

	Speaker				
AM Language	CZ	HU	PL	SK	$\sum$
CZ	53.6	73.8	62.9	45.7	59.0
HU	33.7	28.6	47.1	29.1	34.6
PL	65.0	67.6	46.4	51.1	57.5
RO	53.6	69.1	44.7	43.8	52.8

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 5 for the source language Czech, and in Table 6 for the source language Hungarian. The Hungarian knowledge-based acoustic model significantly outperforms the (Hungarian grapheme-based) baseline, with an 21.1% overall WER. It is worth mentioning that the Czech and Slovakian speaker test sets achieve a surprisingly low 6.4% and 9.1% WER respectively on the Hungarian text test set.

Table 5. WER[%] for Latin-Czech source-target G2P model. Acoustic model training set: 76 hours.

	Latin Test Text			
Speaker	CZ	HU	PL	$\sum$
CZ	43.8	28.2	49.1	40.4
HU	48.7	40.0	58.7	49.1
PL	53.3	18.2	53.2	41.6
SK	30.3	30.0	44.0	34.8
$\sum$	43.9	28.9	50.8	41.2

Table 6. WER[%] for Latin-Hungarian source-target G2P model. Acoustic model training set: 567 hours.

	Latin Test Text			
Speaker	CZ	HU	PL	$\sum$
CZ	19.4	6.4	28.0	17.9
HU	25.0	25.4	20.2	23.5
PL	28.9	15.4	41.3	28.5
SK	20.4	9.1	22.9	17.5
$\sum$	22.6	12.5	28.1	21.1

4.2 USG results

The results for the three-language joint acoustic models are in Table 7. Among the three-language USG models, the Czech-Hungarian-Romanian model had the best performance with a competitive overall 21.9% WER. When adding Romanian, we got the best experimental results of 20.4% with the four-language USG model (see in Table 8). We also measured the WER on any combination of three of the four languages, and found that each language contributed to the four-language model.

It is worth mentioning, that compared to the knowledge-based Hungarian model (Table 6), the results on the Polish speaker test set improved by a significant 6.5% (absolute). This could be due to the ability of the four-language model to generalize better over different speaker test tests. This generalizing ability intensifies when adding training data of a new language, as the models of similar graphemes are merged, and work better on different native language speaker test sets.

Table 7. WER[%] for all the three-language USG models.

	Speaker				
AM Language	CZ	HU	PL	SK	$\sum$
CZ+HU+PL	28.2	28.2	27.7	22.4	26.6
CZ+HU+RO	23.3	21.4	23.9	19.2	21.9
CZ+PL+RO	24.6	33.1	25.6	19.8	25.8
HU+PL+RO	24.8	21.5	25.7	20.7	23.2

Table 8. WER[%] for USG model of Czech, Hungarian, Polish and Romanian (CZ+HU+PL+RO).

	Latin Test Text			
Speaker	CZ	HU	PL	$\sum$
CZ	20.4	11.8	30.7	21.0
HU	21.1	14.6	25.7	20.5
PL	23.0	10.0	33.0	22.0
SK	14.5	12.7	24.8	17.3
$\sum$	19.9	12.2	29.0	20.4

The most striking results in Tables 6 and 8 were that all but the Hungarian speaker test sets performed better on the Hungarian text test set. 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 addition to that, they were pronouncing the named entities using their native pronunciation, which also increased the WER.

Similarly, the results on the Hungarian speaker test set also improved by 3% (absolute) with the four-language USG model compared to the knowledge-based Hungarian model in Table 6. This was supposedly also because the pronunciation of the Hungarian speakers deviated from the one defined in the knowledge-based model, and the four-language model was able to generalize better.

Finally, the results show that the experiments conducted on the Hungarian origin text test set yielded to the best results 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.

4.3 Conclusions

In this paper, we introduced two acoustic modeling techniques for a target language independent medieval Latin speech recognizer to elevate 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 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 building an acoustic model without source language speech data, we presented two approaches: knowledge-based G2P modeling, and USG modeling.

The results showed that both methods outperform by far the best baseline system. We found that the best model was the four-language USG model. When comparing it to the knowledge-based Hungarian phoneme-based model, which was using expert knowledge to map words to phoneme sequences, and trained on larger amount of data, it seemed that the four-language USG model was better in evening out the inconsistencies of the pronunciations in different speaker test sets.

Future research directions include acquiring a considerable amount of medieval speech and textual data, as well as implementing a more refined G2P modeling using a unified phoneme inventory set. Furthermore, adding more data when using the USG approach may result in even higher recognition accuracy, allowing dictational applications.

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)