# Unified Simplified Grapheme Acoustic Modeling for Medieval Latin LVCSR



Lili Szabó, Péter Mihajlik, András Balog, Tibor Fegyó

# What is the problem with Latin speech recognition?

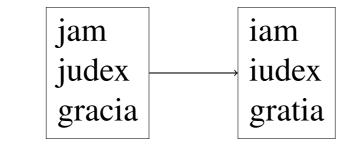
- Latin is not spoken natively
- There is no available speech database, and it is resource-heavy to create one
- Many variants/dialects exists, and we can only make guesses about the pronunciation
- The pronunciation mainly depends on
  - the **era** of the read text
- the **georaphical region** where the text originates from
- the **native language** of the speaker

#### Text data

Regions of origin: Kingdom of Bohemia (CZ), Kingdom of Hungary (HU), Kingdom of Poland (PL)

- In-domain data (Monasterium): medieval charters (HU), 480k/35k token/type
- Background data (Latin Library): historical texts, 1.3M/115k token/type

# Spelling variants



### Speech data

- CZ: 76 hours
- HU: 567 hours (G2P) or 112 hours (grapheme and USG)
- PL: 31 hours
- RO: 35 hours

### Test data

- Independent medieval charters
- Region of read text: CZ, HU, PL
- Native language of test speakers: CZ, HU, PL, SK

Table 1: Perplexity/OOV rate

	Te			
Corpus	CZ	HU	PL	All
Monasterium	551	82	3130	671
Latin Library	3266	3549	2305	4303
Interpolated	924	82	2288	953

# Perplexity measures on test

# Unified Simplified Grapheme (USG) Model

Table 7: Simplification examples for the unified model.

Language	CZ	HU	PL	RO
Orthographic form	řekl	őz	miś	apă
USG transcription	rekl	OZ	mis	apa

Table 8: WER[%] for all the three-language USG models.

Baseline Grapheme Model

Speaker

GR

Table 5: WER[%] for Czech-Latin source-

target G2P model. Acoustic model training

Speaker CZ HU PL \(\sum\_{\text{S}}\)

Latin Test Text

43.8 28.2 49.1 40.4

48.7 40.0 58.7 49.1

53.3 18.2 53.2 41.6

30.3 30.0 44.0 34.8

43.9 28.9 50.8 41.2

set: 76 hours.

CZ

HU

PL

SK

ts

AM Language CZ HU PL SK \(\sum\_{\text{\colored}}\)

Czech, Hungarian, Polish and Romanian (CZ, HU, PL, RO).

53.6 73.8 62.9 45.7 59.0

33.7 28.6 47.1 29.1 **34.6** 

65.0 67.6 46.4 51.1 57.5

53.6 69.1 44.7 43.8 52.8

• Only those grapheme models are retained that are part of the Latin alphabet

Table 2: Word Error Rate (WER[%]) results for monolingual grapheme-based acoustic models of

Knowledge-based grapheme-to-phoneme (G2P) mapping

Table 3: Latin digraph context-insensitive rewrite rules.

CZ e oe f kv

Table 4: Latin context-sensitive rewrite rules. V: vowel, VP: palatal vowel, ^VP: everything but a

palatal vowel, C: consonant, \*: zero or any,  $\hat{}$ : beginning of word,  $\hat{}$  and  $\hat{}$ : not s, t or x.

rule cVP c^VP VC\*ch ^C\*ch guV guC [

Digraph

ae oe ph qu

e ø f kv

gu

gv

gu

gu

tsi

training set: 567 hours.

Speaker

CZ

HU

 $\hat{s}tx$ |tiV|tiC|

Table 6: WER[%] for Hungarian-Latin

source-target G2P model. Acoustic model

Latin Test Text

CZ HU PL \

19.4 **6.4** 28.0 17.9

25.0 25.4 20.2 23.5

28.9 15.4 41.3 28.5

20.4 **9.1** 22.9 17.5

22.6 12.5 28.1 **21.1** 

• All graphemes are trained

CZ

HU

RO

	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
	1				

Table 9: WER[%] for USG model of Czech, Hungarian, Polish and Romanian (CZ+HU+PL+RO).

	<i>,</i>					
		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	

# Language model

- 3-gram language model
- Kneser-Ney smoothing
- Interpolating the two corpora
- SRILM [2]

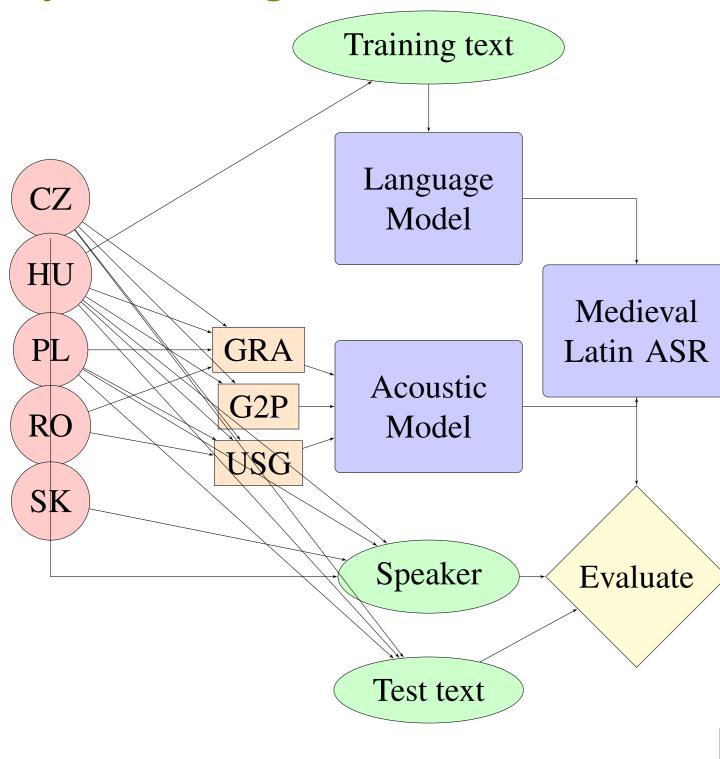
### Dimensions of data

Native language of test speakers: CZ, HU, PL, SK

Region of read text: CZ, HU, PL Speech data: CZ, HU, PL, RO

Model type: baseline, knowledge-based, USG

### System diagram



### Acoustic model

- Mel-Frequency Cepstrum + Energy features were used with Linear Discriminant Analysis (LDA) + Maximum Likelihood Linear Transformation (MLLT), with a splice context of  $\pm 4$  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 [1] activation function.
- Prior to DNN training, a Gaussian 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].

**GRA**: baseline grapheme model **G2P**: grapheme-to-phoneme model **USG**: Unified Simplified Grapheme model

Figure 1: Medieval Latin Speech Recognizer

# Conclusions

- Four-language USG is the best
- It is able to generalize over different speaker test sets

## 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). pp. 901–904 (2002)

