Unified Simplified Grapheme Acoustic Modeling for Medieval Latin LVCSR







Lili Szabó, Péter Mihajlik, András Balog, Tibor Fegyó

lili@speechtex.com



Motivation

• Digitizing medieval charters when optical character recognition in not sufficient

Challenges

- 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

jam		iam
judex		iudex
gracia		gratia

Language model

- 3-gram language model
- Kneser-Ney smoothing

Monasterium 551/11.8

Latin Library | 3266/7.8 3549/1.6

924/3.9

- Interpolating the two corpora
- SRILM [2]

Corpus

Interpolated

Perplexity measures on test

Text region

HU

82/0.9 3130/18.3 479/10.5

2305/5.5 | 2992/9.7

2288/5.5 672/3.5

Table 1: Perplexity/OOV rate (%)

System diagram	1	
	Training text	
CZ	Language Model	
HU PL GRA		Medieval Latin ASR
RO G2P USG	Acoustic Model	
SK	Speaker	Evaluate

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

Figure 1: Medieval Latin Speech Recognizer

Test text

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

Acoustic model

- TODO bullet 1
- TODO bullet 2
- TODO bullet 3
- TODO bullet 4

All

Dimensions of data

- Region of training text: HU, mixed
- Speech data: CZ, HU, PL, RO
- Model type: grapheme, G2P, USG
- Native language of test speakers: CZ, HU, PL, SK
- Region of test text: CZ, HU, PL

Baseline Grapheme Model

- All graphemes are trained
- Only those grapheme models are retained that are part of the Latin alphabet, e.g.
- -keeping model of r
- throwing away model of ř

Table 2: 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				51.1	
RO	53.6	69.1	44.7	43.8	52.8

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

Figure 2: Latin digraph context-insensitive rewrite rules and context-sensitive rewrite rules. V: vowel, VP: palatal vowel, ^VP: everything but a palatal vowel, C: consonant, *: zero or any, ^: beginning of word, $\lceil stx \rceil$: not s, t or x.

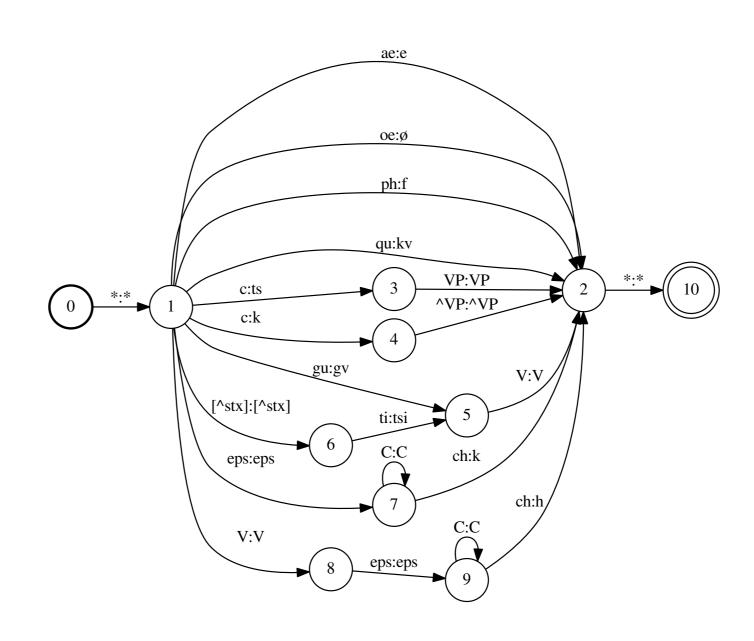


Table 3: WER[%] for Czech-Latin sourcetarget G2P model. Acoustic model training set: 76 hours. Latin Tost Toyt

	Latin Test Text				
Speaker	CZ	HU	PL	\sum	
CZ		28.2			
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 4: WER[%] for Hungarian-Latin source-target G2P model. Acoustic model training set: 567 hours.

	Latin Test Text					
Speaker	CZ	HU	PL	\sum		
CZ		6.4				
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		

Unified Simplified Grapheme (USG) Model

• Utilizing many available language resources in the hopes that statistical variations help generalizing over different pronunciations

Table 5: Simplification examples for the unified model.

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

Table 6: 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

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

тт,	10 11 L 11(0).							
		Latin Test Text						
5	Speaker	CZ	HU	PL	\sum			
($\mathbb{C}\mathbb{Z}$	20.4	11.8	30.7	21.0			
ŀ	HU	21.1	14.6	25.7	20.5			
I	PL	23.0	10.0	33.0	22.0			
S	SK	14.5	12.7	24.8	17.3			
	\sum	19.9	12.2	29.0	20.4			

Conclusions

- Knowledge-based G2P modeling is good, but time consuming and restricted
- Four-language USG modeling 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)