On the use of grapheme models for searching in large spoken archives

Jan Švec[1], Josef V. Psutka[2], Jan Trmal[3], Luboš Šmídl[2], Pavel Ircing[2], Jan Sedmidubsky[4]

[1]NTIS, [2]Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic, [3]Center for Language and Speech Processing, Johns Hopkins University, USA, [4]Faculty of Informatics, Masaryk University, Brno, Czech Republic

Overall description

Our goal is to evaluate the effect of the expert-created pronunciation lexicons used to build phoneme-based models in comparison with the lexicon-free grapheme-based models.

- Spoken term detection (STD) task
- Comparison of grapheme- and phoneme- based acoustic models
- Elimination of grapheme-to-phoneme (G2P) models
  - Especially in the search phase
- USC-SFI MALACH data, difficult for G2P
  - German word *führer*
  - Slovakian town *Kežmarok*
  - Jewish name *Lejerowisz*
- Evaluated on English and Czech
- Introduction of the grapheme-mapped word index

Phoneme-based acoustic models

- Expert-defined pronunciation lexicons
- Difficult to transcribe the query into sequence of phonemes on-the-fly

Grapheme-based acoustic models

- Direct mapping of graphemes to the context-dependent acoustic units
- Exactly one grapheme sequence for each word in the recognition lexicon

Acoustic models

- Kaldi DNN-based training
  - layer-wise RBM pre-training, SGD, sMBR
  - 5 hidden layers (2,048 neurons each), softmax output layer
  - 12-dimensional PLP coefficients (Cepstral Mean Normalized), first and second derivates
- Two sets of acoustics models
  1. the baseline using the phonemes as context-dependent acoustic units with phonetic transcription generated from the pronunciation lexicons
  2. the grapheme-based models using just the graphemes of the lexicon words
- Language model is the same for grapheme- and phoneme-based models

Statistics of development and test sets

	English		Czech	
	Dev	Test	Dev	Test
LVCSR vocabulary	22,723		252,082	
# of graphemes	26		39	
# of phonemes	38		41	
#speakers	10	10	10	10
OOV rate	1.0%	0.7%	3.2%	2.6%
#IV terms	710	735	1762	1764
#OOV terms	154	78	1251	1090
dataset length [hours]	11.1	11.3	20.4	19.4

Recognition error rates

		Dev data		Test data	
		Grphm.	Phnm.	Grphm.	Phnm.
English	words	26.16	25.39	21.21	20.80
	sub-words	23.51	22.15	23.18	21.33
Czech	words	27.66	23.98	23.12	19.11
	sub-words	20.36	19.21	16.51	16.13

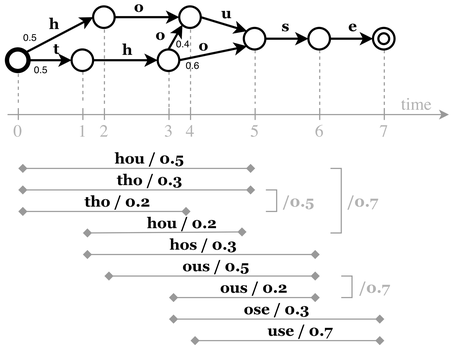
STD task decomposition

- Speech indexing (off-line)
  - Inverted index of sub-word units trigrams
- Putative hits detection
  - Search trough the inverted index and cluster the results
- Term score estimation
  - Use Siamese neural networks to determine the scores

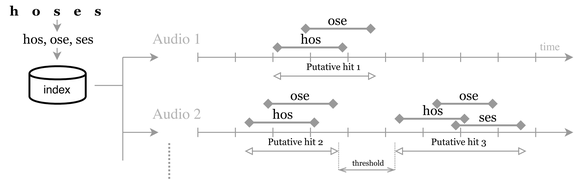
Grapheme-mapped word index

- Speech recognition based on phonemes
  - On both word- and sub-word- levels
- Indexation of grapheme lattices created from word-level lattices
- Putative hits detection uses only graphemes
  - graphemes of the query
  - indexed grapheme lattices
- **Elimination of G2P during the search**

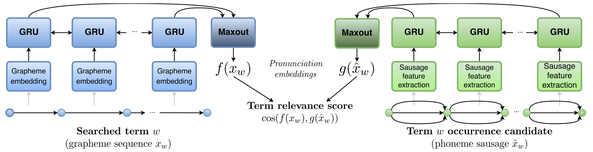
Indexation of grapheme lattices



Searching using sub-word units



Relevance score estimation



$$l(w,\bar{w}) = \frac{1}{2}\cdot\Big[\max\{0, m + d(f(x_w), g(\hat{x}_{\bar{w}})) - d(f(x_w), g(\hat{x}_{\bar{w}}))\} + \max\{0, m + d(f(x_{\bar{w}}), g(\hat{x}_{\bar{w}})) - d(f(x_{\bar{w}}), g(\hat{x}_{\bar{w}}))\}\Big]$$

STD performance on test data

		Dev data		Test data	
	Searched terms	Grphm.	Phnm.	Grphm.	Phnm.
English	IV	0.7759	0.7970	0.6991	0.7447
	OOV sub-word	0.4176	0.3808	0.2677	0.3799
	IV+OOV$_{\text{sub-word}}$	0.6912	0.7070	0.6394	0.7042
	OOV proxy	0.2481	0.2706	0.2105	0.3080
	IV+OOV proxy	0.6804	0.7005	0.6506	0.6992
Czech	IV	0.8227	0.8224	0.8202	0.8277
	OOV sub-word	0.6644	0.6591	0.6777	0.6818
	IV+OOV$_{\text{sub-word}}$	0.7541	0.7546	0.7621	0.7723
	OOV proxy	0.3163	0.4942	0.3353	0.5031
	IV+OOV proxy	0.6125	0.6905	0.6350	0.7090

STD performance using grapheme-mapped word index

		IV	OOV	IV+OOV
English	Graphemes	0.6991	0.2677	0.6394
	+ grph.-mapped index		0.4417	0.6542
	Phonemes	0.7447	0.3799	0.7042
	+ grph.-mapped index		0.3623	0.6895
Czech	Graphemes	0.8202	0.6777	0.7621
	+ grph.-mapped index		0.6260	0.7436
	Phonemes	0.8277	0.6818	0.7723
	+ grph.-mapped index		0.6707	0.7699

Conclusion

- Grapheme- vs. phoneme-based models
  - similar performance for Czech
  - for English the grapheme-based models are slightly worse
- Performance on English improved by using the grapheme-mapped word index
- Complete elimination of the G2P algorithm from the search phase of the STD

References

Josef Psutka et al. (2011). "System for Fast Lexical and Phonetic Spoken Term Detection in a Czech Cultural Heritage Archive". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2011.1, p. 10. ISSN: 1687-4722. DOI: 10.1186/1687-4722-2011-10

Jan Trmal et al. (2017). "The Kaldi OpenKWS System: Improving Low Resource Keyword Search". In: *Proc. Interspeech 2017*, pp. 3597–3601. DOI: 10.21437/Interspeech.2017-601

Jan Švec et al. (2017). "A Relevance Score Estimation for Spoken Term Detection Based on RNN-Generated Pronunciation Embeddings". In: *Proc. Interspeech 2017*, pp. 2934–2938. DOI: 10.21437/Interspeech.2017-1087

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR GBP103/12/G084. Jan Trmal was supported by the NSF grant No CRI-1513128.