Comparison of Slovak and Czech Speech Recognition Based on Grapheme and Phoneme Acoustic Models

Slavomír Lihan, Jozef Juhár, Anton Čižmár

Department of electronics and multimedia telecommunications
Technical University of Košice, Slovakia
slavomir.lihan@tuke.sk, jozef.juhar@tuke.sk, anton.cizmar@tuke.sk

Abstract

Grapheme based mono-, cross- and bilingual speech recognition of Czech and Slovak is presented in the paper. The training and testing procedures follow the MASPER initiative that was formed as a part of the COST 278 Action. All experiments were performed using Czech and Slovak SpeechDat-E databases. Grapheme-based models gave equivalent recognition performance compared to phoneme-based models in monolingual as well as bilingual case. Moreover bilingual SK-CZ speech recognition is equivalent to monolingual recognition, which indicates the possibility to share Czech and Slovak speech data for training bilingual grapheme-based acoustic models usable for recognition of Slovak as well as Czech. Also the promising results confirmed the presumption, that languages with a close grapheme-to-phoneme relation are well suited for grapheme-based speech recognition.

Index Terms: grapheme unit, bilingual recognition, SpeechDat

1. Introduction

Phonemes and allophones serve as basic speech units for training of HMM based acoustic models in most of today's speech recognizers. Very important part of each speech database is a dictionary, which can contain thousands of phonetically transcribed words. Phonetic transcription of lexicon words is a time consuming task. Although this task can be automated with the help of rules covering phonetic properties of particular language, because of lots of exceptions a final manual check must be made.

In [1] mapping the orthographic transcription of words directly onto HMM state models using phonetically motivated decision questions was proposed.

In [2] grapheme-based acoustic sub-word units together with automatic generation of questions for decision tree state-tying to multilingual acoustic modeling was applied. This reduced the effort to find a common set of acoustic sub-word units which in the case of phonemes requires expert phonetic knowledge.

In [3] several decision tree based clustering procedures are performed and compared in order to develop context dependent grapheme based speech recognizers in three different languages and multilingual grapheme based recognizers were designed. In [4] a grapheme based speech recognition system for Russian was investigated.

In [5] a continuous speech recognizer which uses both phoneme and grapheme as subword units has been investigated. It has been shown that ASR using just grapheme as subword unit yields acceptable performance, which could be further improved by introducing phonetic knowledge in it.

Experiments on different languages have shown that the quality of the resulting recognizer significantly depends on the grapheme-to-phoneme relation of the underlying language [2], [3], and [4]. Since Slovak and Czech are languages with a fairly close grapheme-to-phoneme relation they should be very well suited to be candidates for such an approach.

In [6] a crosslingual and bilingual speech recognition with context dependent phoneme-based acoustic models trained on SpeechDat-E databases for Czech and Slovak language was presented. In [7] a grapheme based crosslingual speech recognition carried out within the MASPER initiative was introduced.

In this paper we present our results on creating a grapheme based monolingual, crosslingual and bilingual Slovak and Czech recognizer trained on the SpeechDat-E corpus. We compare the performance of the resulting system to a phoneme based mono/cross/bilingual recognition system that was trained in the course of the COST 278 Action [6], [7].

2. Orthography of Slovak and Czech

Slovak (SK) and Czech (CZ) belong to the family of Slavic languages and have lots of similar features. Both languages use the Roman alphabet for its written communications. However, because they, unlike English, use the rule "write as you hear", the 26 characters of standard Roman alphabet are not enough to represent every phoneme. This problem is overcome by:

- The use of digraphs to represent a single phoneme (**dz**, **dž**, and **ch**).
- The use of diacritic marks (ˇ, ´, ¨, ˆ, °). The most frequent are an acute accent (´) indicating vowels which pronunciation are relatively protracted and a hook (ˇ) over a consonant, meaning that the consonant is palatalized (softened).

The letters **q**, **x** and **w** are only used in foreign words, never in native Slovak or Czech words.

2.1. Slovak grapheme set

Slovak basic grapheme set consists of 46 elements including **w** and **q**. Here are some special observable features of the Slovak language:

- The letters **l** and **r** can function either as a vowel or as a consonant. When functioning as a vowel, they can be long (**ĺ**, **ŕ**) or short.

- The mark (ˇ) indicates that a consonant is ***soft***. If it is not present, the consonant is considered to be hard. The consonants **d**, **t**, **n**, and **l**, however, are made implicitly soft if followed by **i**, **í** or **e**. So, for example, the **t** in the word "**tehla**" (brick) is implicitly pronounced as **ť** ("**ťehla"**). However there are a number of exceptions to the implicit softness, such as in the words "**teraz"** and "**teda"**. This is troublesome when making scripts for automatic phonetic transcription.
- While every Slovak vowel can be either long or short, not all Slovak consonants have a soft counterpart. Here are the soft ones: **č, ď, ľ, ň, š, ť, ž, dž**.
- A consonant standing before the vowel **y** is never soft. Note that **y** is pronounced the same as **i** and stands always as a vowel, not a consonant (for example as compared to "yellow" in English).
- Couples of graphemes **ia, ie**, **iu** are considered as diphthongs. The grapheme **ô** represents diphthong strongly resembling the coincident **uo**.
- Umlaut is used only over the letter **a** (**ä**), but in nowadays colloquial Slovak is almost always pronounced as **e**.
- Five Slovak graphemes are not used in the Czech language: **ô, ä, ľ, Í, ŕ**.

2.2. Czech grapheme set

The basic Czech grapheme set consists of 45 elements including **w** and **q**. Special features of the Czech language are:

- The vowels **i**, **í**, **ě** cause the softening of the preceding consonant **d**, **t**, **n**, which is behaviour the same as in Slovak. Soft version of **l** (**ľ**) does not exist in Czech.
- The vowel **ě** is pronounced as **je** after **p**, **b** and **v** (**pět** = **pjet**), and combination **mě** is pronounced as **mňe**.
- Vowel **y** is the same as in Slovak.
- The long vowels **ú** and **ů** are pronounced the same. The only difference between **ú** and **ů** is their position in the word; **ú** is only used at the beginning of a word, while **ů** is used inside or at the end of a word, but there are some exceptions too.
- Consonant **ř** occurs nowhere else but in Czech. It is rough approximation is **r** with simultaneous **sh** or **zh**.
- Three Czech graphemes are not used in the Slovak language: **ů, ě, ř**.

3. Monolingual recognition experiments

3.1. Training and testing setup

Training process, following REFREC/MASPER procedure [8] [9], begins with flat started context independent grapheme models and ends with context dependent models - trigraphemes. Decision-tree based state clustering was applied on the context dependent models. A broad classification of graphemes was created based on the phonetic properties of phonemes using phoneme - grapheme mapping. The number of mixtures was increased up to 32. Two non-speech sounds - filled pauses (**fil**), speaker noise (**spk**) - were modeled in addition to the graphemes, silence (**sil**) and short pause (**sp**). For both languages digraphs **ch**, **dz** and **dž** were considered and modeled as single graphemes.

Acoustic models performance was evaluated by the WER parameter on standard six different test scenarios, as specified in SpeechDat-E [10]: applications words (A), isolated digits (I), yes/no answers (Q), own names (O), phonetically rich words (W), connected digits (BC).

3.2. Slovak grapheme acoustic models

Set of 45 graphemes was modeled reduced by the grapheme **w** mapped to **v**. Basic training statistics for Slovak grapheme based acoustic models is in Table 1 compared to phoneme based training statistics as presented in [6]. Phoneme acoustic models were created for the set of 51 phonemes and allophones. Table 1 shows that the number of trigraphemes is smaller than the number of triphones. The state tying is more effective for phonemes.

Table 1: *Training statistics for Slovak grapheme and phoneme based acoustic models*

Number of	Graph.	Phone.
Sessions (speakers)	800	800
Training utterances	32845	32 855
Lexicon words	14 907	14 907
Lexicon pronunciations	14 907	14 909
Trigraph./Triphon. in training set	8 322	9 121
Trigraph./Triphon. in lexicon	8 606	9 417
States before clustering	25 026	27 390
States after clustering	3 714	3 666
Clustering reduction	14.8%	13.4%

Speech recognition performance for both sets can be found in Table 2. Context dependent models with tied states and 16 mixtures per state are compared, as they gave better results than the 32-mixture models for both cases. We can see that grapheme-based recognition evidently surpasses phoneme-based one in tests with medium vocabulary – directory assistance names (O) and phonetically rich words (W). In average the grapheme-based recognition seems to be slightly better too.

Table 2: *Comparison of Slovak grapheme and phoneme based speech recognition performance (monolingual)*

	WER (Word Error Rate) [%]						
	A	I	Q	O	W	BC	avg.
Graph.	0.77	0.54	0.00	7.55	8.67	1.39	3.15
Phon.	0.43	0.54	0.00	8.16	10.46	1.32	3.49

Progress of average WER evaluated on consequently retrained models with an increasing number of mixtures is shown on Fig. 1. Number attached to the model name represents mixtures count per state. The results confirmed the assumption that context dependent modeling can cover changes in grapheme pronunciation due to, for example, voicing assimilation in group of consonants. Grapheme models which gave worse performance in context independent mode became better after they were trained in context dependent mode.

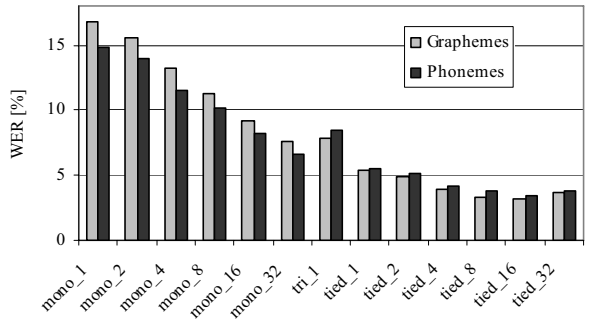

Figure 1: *Progress of performance for consequently trained models*

3.3. Czech grapheme acoustic models

Only 41 graphemes were used – the graphemes **w, ů** and **q** were mapped into the more frequent alternatives **v**, **ú** and pair **k v**. Number of modeled phoneme models was 42 [6]. Training statistics can be found on Table 3. This analysis showed that there are small differences in the training statistics between both Czech model sets.

Table 3: *Training statistics for Czech grapheme and phoneme based acoustic models*

Number of	Graph.	Phone.
Sessions (speakers)	800	800
Training utterances	37164	36 449
Lexicon words	19 313	19 313
Lexicon pronunciations	19 313	20 114
Trigraph./Triphon. in training set	9 650	9 675
Trigraph./Triphon. in lexicon	9 958	10 102
States before clustering	29 010	29 031
States after clustering	4 341	4 254
Clustering reduction	15.0%	14.7%

Recognition performance of grapheme models is again compared to performance of phoneme models and can be found in Table 4. In this case models with 32 mixtures per state are compared, as they gave the best results on average. Czech grapheme models gave noticeable better performance in every particular test.

Table 4: *Comparison of Czech grapheme and phoneme based speech recognition performance (monolingual)*

	WER (Word Error Rate) [%]						
	A	I	Q	O	W	BC	avg.
Graph.	0.55	0.56	0.00	6.96	7.02	1.76	2.81
Phone.	0.94	2.55	0.57	8.13	7.18	1.92	3.55

4. Crosslingual SK-CZ recognition experiments

In the case of graphemes, the same amount of speech material is used to model less diverse acoustic models. Graphemes model a wider part of the acoustic-phonetic space. In general, broader acoustic models should be better suited for crosslingual speech recognition, having shown to be more tolerant to inaccuracies incorporated by the mapping procedure from the source to the target language [5].

Both Slovak and Czech were used as target and source and the obtained results are compared to results with phoneme acoustic models presented in [6]. Target to source grapheme mapping was done using only expert knowledge, because there are only a few differences between both languages. Context dependent models with tied states were used as source. Unseen contexts from target dictionary were tied to source models based on decision tree generated during source acoustic models training.

4.1. Czech as the target language

Only three Czech graphemes which do not exist in Slovak were mapped to similar equivalents: **ů** to **ú**, **ě** to pair **j e**, and **ř** to pair **r ž**. Crosslingual recognition performance for this situation is presented in Table 5.

Table 5: *Czech crosslingual speech recognition with Slovak source acoustic models*

CZ as the target	WER (Word Error Rate) [%]						
	A	I	Q	O	W	B	avg.
Graph.	1.56	9.50	11.60	12.03	14.76	8.01	9.58
Phone.	2.11	7.26	1.25	12.22	15.08	8.90	7.80

As it can be seen from the Table 5, the grapheme models are slight better in almost every particular test except I and Q tests. Surprising is the poor result of the simplest test with only two-word vocabulary (Q). In this case the high value of WER is caused by substitution of word "**ne**" by word "**ano**". Opposite substitution was not seen. This is probably caused by hard pronunciation of the Czech "**ne**" ("no" in English) as opposed to Slovak, where the hard pronunciation of "**ne**" is rare and can be found almost exclusively at the end of words.

4.2. Slovak as the target language

Five Slovak graphemes had to be replaced by similar equivalents in Czech language: **ô** by the pair of **u o**, **ä** by **e**, soft **ľ** by **l**, long **Í** by pair **l l**, and long **ŕ** by pair **r r**. Obtained results are presented in Table 6.

Table 6: *Slovak crosslingual speech recognition with Czech source acoustic models*

SK as the target	WER (Word Error Rate) [%]						
	A	I	Q	O	W	B	avg.
Graph.	2.88	9.09	0.00	9.81	20.30	9.14	8.54
Phone.	2.51	8.28	0.33	8.97	17.93	6.50	7.42

In this case the grapheme models are slight worse in all particular tests (comparing to phoneme models). As expected, phonetically rich words gave the worst performance due to rich occurrence of mapped graphemes. Crosslingual SK-CZ speech recognition shows generally almost triple WER comparing to

monolingual case in both cases, grapheme as well phoneme ones.

5. Bilingual recognition

In [6] bilingual speech recognition system for Slovak and Czech using phoneme acoustic models was presented. The same task was performed considering graphemes as the basic speech units. A common set of 45 graphemes was used - four graphemes were substituted by alternatives as in the monolingual case: **w** by **v**, **ä** by **e**, **ů** by **ú**, **q** by **k v**. Common SAMPA set for phoneme acoustic training contained 55 units [6]. Important training statistics are compared in Table 7.

Table 7: *Comparison of training statistics for bilingual CZ - SK grapheme and phoneme based recognition*

Number of	Graph.	Phon.
Trigraph./Triphon. in training set	11 979	13579
Trigraph./Triphon. in lexicon	12 363	14119
HMM states before clustering	36 000	40761
HMM states after clustering	6 166	6313
Clustering reduction	17.1%	15.5%

The speech recognition performance for bilingual acoustic models was evaluated separately for both languages. Performance obtained with grapheme bilingual models for both languages is compared to performance of phoneme bilingual models with the same training and testing setup (models with 16 mixtures). Comparison can be found in Table 8.

Table 8: *Recognition performance for bilingual CZ - SK grapheme and phoneme recognition*

Test	Slovak test set		Czech test set	
	graph.	phon.	graph.	phon.
A	0.69	0.43	0.43	0.34
I	0.54	0.54	3.57	3.06
Q	0.00	0.00	0.00	0.00
O	6.51	6.94	9.11	8.69
W	10.33	10.71	6.52	6.91
BC	1.43	1.81	2.21	1.94
avg.	3.25	3.41	3.64	3.49

Compared to phoneme-based models the results show that bilingual grapheme models give equivalent recognition performance. Moreover bilingual SK-CZ speech recognition is equivalent to monolingual one, which indicates the possibility to share Czech and Slovak speech data for training bilingual acoustic models usable for the recognition of Slovak as well as Czech language.

6. Conclusions

A set of grapheme-based mono, cross and bilingual recognition tests of Czech and Slovak showed promising results, confirming the presumption, that language with a close grapheme-to-phoneme relation are well suited to this approach. Grapheme-based models gave equivalent recognition performance compared to phoneme-based models in monolingual as well as the bilingual case. Moreover bilingual SK-CZ speech recognition is equivalent to monolingual one, which indicates the possibility to share Czech and Slovak speech data for training bilingual grapheme-based acoustic models usable for recognition of Slovak as well as Czech language.

The worst results were obtained in the crosslingual case indicating an area for further improvement. In our future work we will concentrate on investigating of new methods of tree based clustering and state tying. The other method we envisage is to investigate the possibility of introducing phonetic knowledge in grapheme-based acoustic models, as it was outlined in [7].

7. Acknowledgements

This work was supported by the Slovak Grant Agency VEGA under grant No. 1/1057/04 and the Ministry of Education of the Slovak Republic under research project No. 2003 SP 20 028 01 03. The SpeechDat-E databases were available thanks to collaboration with the Institute of Informatics Slovak Academy of Science. The COST 278 participants in MASPER SIG are acknowledged for their inspiration in relation to the presented work.

8. References

[1] Kanthak, S., Ney, H., "*Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition*", In Proceedings ICASSP 2002, Orlando, Florida, 2002, pp. 845-848.

[2] Kanthak, S., Ney, H., "*Multilingual Acoustic Modeling Using Graphems*", In Proceedings of Eurospeech 2003, Geneva, Switzerland, September 2003, vol. 2, pp. 1145-1148.

[3] Killer, M., Stüker, S., Schultz, T., *"Grapheme based speech recognition"*, In Proc. Eurospeech 2003, Geneva, Switzerland, September 2003, pp. 3141-3144.

[4] Stüker, S., Schultz, T., "*A grapheme based speech recognition system for Russian*", In Proc. SPECOM-2004, Saint-Petersburg, Russia, Sept. 2004, pp. 297-303.

[5] Magimai-Doss, M., Bengio, S., Bourlard, H., "*Joint decoding for phoneme-grapheme continuous speech recognition*", In Proc. ICASSP 2004, Montreal, Canada.

[6] Lihan, S., Juhár, J., Čižmár, A., "*Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases",* In Proc. Interspeech 2005, Lisabon, Portugal, September 2005, pp. 225 – 228.

[7] Žgank, A., Kačič, Z., Lihan, S., Juhár, J., Diehl, F., Vicsi, K., Szaszak, G., "*Graphemes as basic units for crosslingual speech recognition",* In Proc. ITRW ASIDE 2005, November 2005, Aalborg, Denmark, pp. 23-27.

[8] Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G., *"A noise robust multilingual reference recognizer based on SpeechDat (II)"*, In Proc. ICSLP 2000, Beijing, China.

[9] Žgank, A., Kačič, Z., Lihan, S., Juhár, J., Diehl, F., Vicsi, K., Szaszak, G., "*Crosslingual transfer of source acoustic models to two different target languages"*, In Proc. ITRW Robust 2004, Norwich, UK, August 2004.

[10] Heuvel, H. et al., "*Five Eastern European Speech Databases for Voice-Operated Teleservices Completed*", In Proc. Eurospeech 2001, Aalborg, Denmark, 2001.