Comparison of Grapheme and Phoneme Based Acoustic Modeling in LVCSR Task in Slovak

Michal Mirilovič, Jozef Juhár, and Anton Čižmár

Department of Electronics and Multimedia Communication,
Technical University of Košice
{michal.mirilovic,jozef.juhar,anton.cizmar}@tuke.sk
http://kemt.fei.tuke.sk/

Abstract. Phonemes and allophones are the basic speech units for acoustic modeling in the majority of contemporary HMM based speech recognizers. Grapheme-based acoustic sub-word units were applied to multi-lingual and cross-lingual acoustic modeling in many tasks. Grapheme and phoneme based mono-, cross- and bilingual speech recognition of Czech and Slovak in the small and medium vocabulary task has been studied in our previous work. In this article we compare grapheme and phoneme based approach to acoustic modeling and model unit selection in large vocabulary continuous speech recognition (LVCSR) task in Slovak. The main goal of our experimental work is to investigate a possibility to select an optimal set of sub-word units for Slovak LVCSR system.

1 Introduction

In general, there are two common approaches to acoustic modeling for automatic speech recognition. The most frequent approach is based on *phonemes*, where each HMM represents different phoneme or phone. We distinguish *context dependent* and *context independent* models. Especially for LVCSR one of the most complicated tasks is phonetic transcription of lexicon words, which is hardly fully automatized because of many exceptions requiring partial expert approach.

Another approach, that occurs in the last years and issues from tight relation between orthographic and orthoepic description, is based on *graphemes*. This approach was applied for instance in hybrid phoneme-grapheme systems [1] [2] [3], multilingual speech recognition [4] [5], porting acoustic models from one language to another [6], building ASR systems for *under-resourced* languages [7] [8] [9], building ASR systems for languages with close grapheme-to-phoneme relation [10] and other [11] [12].

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. Slovak language has a fairly close grapheme-to-phoneme relation. It should be well suited for grapheme based approach to acoustic modeling. In [13] a speech recognizer which used both phoneme and grapheme as sub-word units has been investigated. It has been shown that ASR using just

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 242–247, 2009.

grapheme as sub-word unit yields acceptable performance, which could be further improved by introducing phonetic knowledge in it.

We applied the phoneme and grapheme based approach to acoustic model training also in LVCSR task [14]. This preliminary comparison of phoneme and grapheme based acoustic models in LVCSR system gave us encouraging results. Grapheme based acoustic models were better in tests with higher order n-gram (bigram and trigram) language models (LM). These results were the reason for our continuing research in this area. Optimizing procedure of unit selection described in this paper is managed by pronunciation effects occurring in Slovak language (assimilation, palatalization, ...) and by their influence on word error rate and other errors of LVCSR system.

2 Basics of Slovak Orthoepy and Orthography

In basic Slovak orthoepy and orthography we have 46 graphemes and 53 phonemes. According to SAMPA phonetic alphabet Slovak phoneme set contains:

- 15 vowels and diphtongs
- 38 consonants (17 sonorants, 8 fricatives, 9 plosives, 4 affricates)

Tight relation between orthographic and orthoepic representation of speech is *typical* for Slovak language. The main differences between this two representations are in:

- *Softening*, where some Slovak consonants have a soft counterpart (`c-č`, `d-ď`, `l-ľ`, `n-ň`, `s-š`, `t-ť`, `z-ž`, `dz-dž`) and in many cases hard consonants `d,t,n,l` are pronounced softly, depending on context (if they are followed by short and long form of `i` and `e`). Actually, there are rules with many exceptions.
- *Voice / unvoice assimilation*, where voiced consonants are pronounced as unvoiced if they are followed by unvoiced consonant and unvoiced consonants are pronounced as voiced if they are followed by voiced consonant. There are also rules with many exceptions.
- *Existence of two different graphemes with the same pronunciation*, where pairs: vowels `i` and `y`, liquids `v` and `w` and vowel `ä` (wide a) and `e` (in contemporary Slovak) have the same pronunciation.
- *Existence of two graphemes represent one phoneme and reversely*, where digraphemes `dz`, `dž`, `ch` represent single phonemes, single grapheme `ô` represents diphtong `u o`, single grapheme `q` represents phonemes `k v` and single grapheme `x` represents phonemes `k s`.

3 Experimental Setup

3.1 Levels of Phonetic Transcription

Based on these differences we made 5 levels of phonetic transcription:

1. *Zero level phonetic transcription.* At this level no phonetic transcription was applied.
2. *Basic level of phonetic transcription.* The following basic transcription rules were applied:
   - considering each digrapheme as one grapheme
   - considering `y/i` like one grapheme
   - considering `q` like two graphemes `k v`
3. *Middle level of phonetic transcription.* All rules from the basic level supplemented with the following softening rules were applied:
   - softening `d`, `t`, `n`, and `l` before `e/i/í` with exceptions
4. *Advanced level of phonetic transcription.* The assimilation rules were added to the middle level:
   - voiced (`b d ď dz dž g h z ž v`)/unvoiced (`p t ť c č k ch s š f`) assimilation
   - unvoiced (`p t ť c č k ch s š f`)/voiced (`b d ď dz dž g h z ž v`) assimilation
5. *Full phonetic transcription.* At this level all transcription rules were applied.

We started with grapheme based transcription and then we continuously added some transcription rules. Final level of transcription was full phonetic transcription (see procedure above).

3.2 Acoustic Modeling

Acoustic models were trained on SpeechDat-SK corpus. Training algorithms and procedures were implemented by the HTK tools and can be considered as standard. They come out from [15] and results to set of speaker independent cross-words triphones with 16 tied states. The training procedure consists of these steps:

1. *Embedded training* of context independent (CI) models (monophones). Result of this process were CI HMMs with increased Gaussian mixtures up to 32. Training was performed on acoustic data with word level transcription.
2. *Alignment* of phoneme borders. Result of this process were acoustic data with phone level transcription. Recognition was performed with CI HMMs with 32 Gaussian mixtures (step above).
3. *Isolated training* of CI models (monophones). Result of this process were new CI HMMs (without increase of Gaussian mixtures). Training was performed on acoustic data with new phone level transcription.
4. *Cross-word* context dependent (CD) models (triphones). Result of this process were new CD HMMs (without increase of Gaussian mixtures). These HMMs was made by the copying of previous CI HMMs and reestimation. Each CD HMM was copied from CI HMM according to its central phoneme.
5. *Decision tree* building and *tying* states. Result of this process were CD HMMs with tied states. Decision tree was build on a base of hand created phone/grapheme classes [13].
6. Increasing of *Gaussian mixtures* (16). Result of this final step were tied-state crossword CD HMMs with 16 Gaussian mixtures.

The SpeechDat family of corpora are more suitable for training and testing from small to medium vocabulary speech recognition systems. Nevertheless, each session (speaker) of the SpeechDat family also contains utterances: so-called *phonetically rich sentences*, which are usually used to initialize HMMs in the training process. Each utterance can be occurred in the whole corpus approximately from 1 to 10 times (never 2 times per same speaker). These utterances we used for LVCSR testing.

In case of LVCSR systems this option could be considered as not standard. In our case we used it because of absence of a more convenient speech corpus with sufficient amount of acoustic-phonetic data.

3.3 Language Modeling

Language model was trained on 2 corpora. The first one, Slovak National Corpus (SNC) [16] is an electronic database of Slovak linguistic resources, containing wide spectrum of language styles, genres, subject domains and additional linguistic information. The corpus version *prim-3.0* contains more than 350 millions words. It is publicly available since January 2007. It was built for linguistic purposes primarily. Despite of this, due to lack of other suitable linguistic data we used the SNC as a basic training corpus.

The second one, TUKE Text Corpus we are systematically collecting from Internet text resources in Slovak language by using a software tool developed for the purpose. The tool is written in Java and it is based on the RSS principle. It uses approximately 250 RSS channels and on average it collects 2.8 MB of text data per day. At the time of the experiments performing approximately 55 millions words were collected.

Both corpora were coupled together and used in LM training. Result of training process was bigram LM with 240k of unique words, smoothed by Knesser-Ney smoothing method. Lexicon words were selected according their occurrence in the training data.

4 Testing and Results

In the testing part of the experiment we used 1139 utterances (phonetically rich sentences) spoken by 200 speakers from the testing part of the SpeechDat-SK corpus. Speech recognition was performed by HTK tool HDecode with *maximum number of active models* set to 3072, *word end pruning* enabled and set to 50, *word insertion probability* set to -8.0, *grammar scale factor* set to 12 and *beam searching* enabled and set to 400.

Word error rate (WER) was computed between reference and hypothesis as:

$$WER = \frac{SUB + INS + DEL}{N} \tag{1}$$

where SUB is error by the substitution, INS is error by the insertion, DEL is error by the deletion and N is number of the words in the reference.

Results of the experiment are presented in the Fig. 1. The best result was reached on the HMM set with the middle level of phonetic transcription (30.786%).

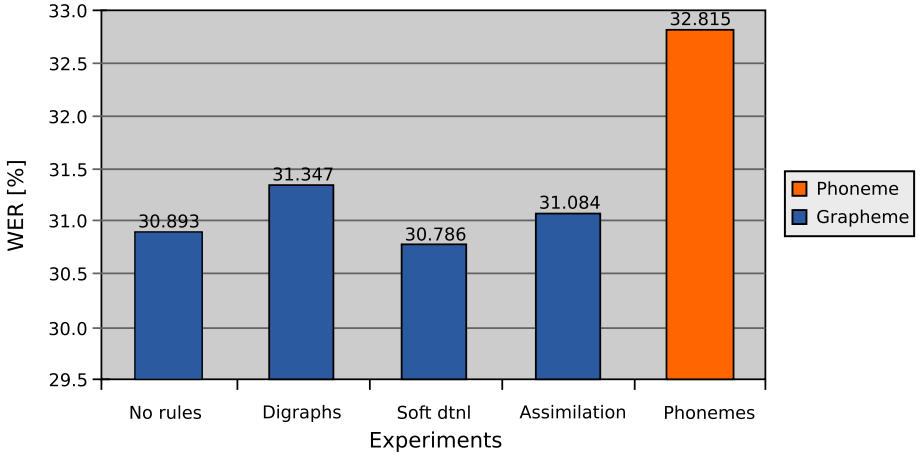

Fig. 1. Results

5 Conclusion

Experiments showed that graphemes-based LVCSR outperformed phoneme-based one. In all cases the error was *statistically spread*, so we did not see any regularity. It means, that linguistic information in language model was generally more important that acoustic one, coming from phonetic transcription. So far, we did not study this phenomenon more deeply, because our LVCSR system is *under-resourced* yet.

In the near future we need to continue in collecting acoustic and linguistic resources and working on better language model (e.g. morpheme based). Besides this we will study more deeply phenomenon of grapheme based speech recognition to use it for improving of our future ASR systems.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-0369-07 and by the Slovak Ministry of Education under contracts No. AV 4/0006/07 and AV 4/2016/08.

References

1. Schukat-Talamazzini, E.G., Niemann, H., Eckert, W., Kuhn, T., Rieck, S.: Automatic speech recognition without phonemes. In: Proceeding of the Eurospeech, Berlin, September 22-25, pp. 129–132 (1993)
2. Magimai-Doss, M., Stephenson, T.A., Bourlard, H., Bengio, S.: Phoneme-grapheme based speech recognition system. In: Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands, November 30 - December 4, pp. 94–98 (2003)

3. Magimai-Doss, M., Bengio, S., Bourlard, H.: Joint decoding for phoneme-grapheme continuous speech recognition. In: Proceedings of ICASSP, Quebec, Kanada, May 17-21, pp. 177–180 (2004)
4. Kanthak, S., Ney, H.: Multilingual acoustic modeling using graphemes. In: Proceeding of the Eurospeech, Geneva, Switzerland, September 1-4, pp. 1145–1148 (2003)
5. Killer, M., Stüker, S., Schultz, T.: Grapheme based speech recognition. In: Proceeding of the Eurospeech, Geneva, Switzerland, September 1-4, pp. 3141–3144 (2003)
6. Schultz, T.: Towards rapid language portability of speech processing systems. In: Proceedings of the Conference on Speech and Language Systems for Human Communication, SPLASH 2004, Delhi, India, November 17-19 (2004)
7. Rubagotti, E.: Is it possible to train a speech recognition system on text only? In: Interspeech 2006 - ICSLP, Stellenbosch, South Africa, April 9-11 (2006)
8. Le, V.B., Besacier, L.: Comparison of acoustic modeling techniques for vietnamese and khmer asr. In: Interspeech 2006 - ICSLP, Pittsburgh, USA, September 17-21, pp. 129–132 (2006)
9. Charoenpornsawat, P., Hewavitharana, S., Schultz, T.: Thai grapheme-based speech recognition. In: Proc. of the HLT-NAACL, New York City, USA, June 5-7, pp. 17–20 (2006)
10. Stüker, S., Schultz, T.: A grapheme based speech recognition system for Russian. In: Proceedings of SPECOM 2004, Petersburgh, Russia, September 20-22 (2004)
11. Kanthak, S., Ney, H.: Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In: Proceeding of the ICASSP, Orlando, Florida, May 13-17, pp. 845–848 (2002)
12. Schillo, C., Fink, G.A., Kummert, F.: Grapheme based speech recognition for large vocabularies. In: Proceeding of the ICSLP, Beijing, China, October 16-20, pp. 584–587 (2000)
13. Lihan, S., Juhár, J., Čižmár, A.: Comparison of Slovak and Czech speech recognition based on grapheme and phoneme acoustic models. In: Interspeech 2006 - ICSLP, Pittsburgh, USA, September 17-21, pp. 149–152 (2006)
14. Mirilovič, M., Juhár, J., Čižmár, A.: Large vocabulary continuous speech recognition in slovak. In: Proc. Int. Conf. on Applied Electrical Engineering and Informatics - AEI 2008, Greece, September 8-11 (2008)
15. Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G.: A noise robust multilingual reference recogniser based on SpeechDat(II). In: Proc. ICSLP 2000, Beijing, China, October 16-20, vol. 3, pp. 370–373 (2000)
16. Šimková, M.: Slovak national corpus history and current situation. In: Insight into the Slovak and Czech Corpus Linguistics, Veda, Bratislava, pp. 151–159 (2006)