Conversational telephone speech recognition for Lithuanian

Rasa Lileikyte, Lori Lamel, Jean-Luc Gauvain

CNRS/LIMSI, Spoken Language Processing Group, 91405 Orsay Cedex, France
lileikyte@limsi.fr, lamel@limsi.fr, gauvain@limsi.fr

Abstract. The paper presents a conversational telephone speech (CTS) recognition system for the low-resourced Lithuanian language, developed in the context of IARPA-Babel program. We compare phoneme-based systems and grapheme-based systems to establish whether or not it is necessary to use a phonemic lexicon. We explore the impact using the additional Webdata for language modeling and additional untranscribed data for semi-supervised training. The experiments are performed for two conditions: Full Language Pack (FLP) and Very Limited Language Pack (VLLP). Graph-based systems are shown to give comparable results to phoneme-based ones. Adding Web texts improves the performance of both the FLP and VLLP system. The best VLLP results are achieved using both Web texts and semi-supervised training.

Keywords: conversational telephone speech, Lithuanian, KWS, STT

1 Introduction

The Lithuanian language is one of the least spoken European languages, with only about 3.5 million speakers. Lithuanian belongs to the Baltic subgroup of Indo-European languages, and it is the oldest surviving Indo-European language. The language was standardized during the late 19th century and the early 20th century. Having preserved the majority of phonetical and morphological features [19], Lithuanian has rich inflection, complex stress system, and flexible word order. It is based on the Latin alphabet and has some additional original characters, also characters borrowed from other languages. There are two main dialects - Aukštaitian (High Lithuanian), and Samogitian (Žemaitian or Low Lithuanian), each with sub-dialects. The dominant dialect is Aukštaitian, spoken in the east and middle Lithuania by 3 millions speakers. Samogitian spoken in the west of the country by about 0.5 millions speakers.

This paper presents the development of conversational telephone speech (CTS) recognition system for Lithuanian language. Todays speech recognition systems make use of statistical models and are typically trained on large data sets. Three main resources are needed: 1) telephone speech recordings with corresponding transcriptions for acoustic models, 2) written text for language modeling, 3) and a pronunciation dictionary. There have been few studies reporting on speech recognition for Lithuanian, in part due to sparcity of linguistic resources. In

[18] a Lithuanian broadcast speech recognition system is described trained on only 10 hours of transcribed speech. An uni-code based graphemic system described in [4] reports on the transcription of conversational telephone speech in Lithuanian.

Our system was built in context of IARPA Babel project, as for [4] provided training resources for two conditions: full language pack (FLP) with approximately 40 hours of transcribed telephone speech and very limited language pack (VLLP) comprising about 3 hours of transcribed data. About 40 hours of untranscribed speech was also available to produce the transcriptions in semi-supervised manner. Transcribing conversational telephone speech is more complex task than transcribing broadcast news. Spontaneous telephone speech has a high variability of speaking rates, styles, grammar rules are not strictly followed, also the impact of limited bandwidth, distorted audio channels. Additional text corpora was also provided for training. We used data prepared by our partner BBN, which contains texts collected from the Web such as Wikipedia, subtitles, and other. The text corpora consists of 26M words. However, 40 hours of transcribed data and 26M text words is a very small amount compared with the 2000 hours of transcribed audio and over a billion words of language modeling text, that are available for the English language [17].

The pronunciation dictionary is an important component of the system. To generate one, a grapheme based or phoneme based approach can be used. The advantage of using graphemes is that models can be easily defined. Grapheme based systems have been shown work well for various languages [12], [14]. Yet, some languages as English have a weak correspondence between graphemes and phonemes, and using graphemes leads to system performance degradation. Phoneme based systems usually provide good results as they have a stronger correlation with the audio. However, designing the rules requires the linguistic skills of expert, making it an expensive process. The Lithuanian language has a quite strong dependency between the orthographic transcription and the phonetic form. Conversion rules are implemented easily, compared with the English language that requires numerous exceptions. In our study we developed CTS system for the Lithuanian language and addressed the following questions: 1) is a phoneme-based system better than grapheme-based one, 2) how can additional resources (untranscribed audio and web texts) be used to improve the system performance, and their impact with respect to the original training conditions.

We describe the phonemic inventory of the Lithuanian language in Section 2. In Section 3 the experimental setup is defined. In Section 4 the results are presented for different sets of graphemes and phonemes, also the results when semi-supervised training and Web text are used. Finally, in Section 5 we make conclusions.

2 Lithuanian phonetic inventory

The Lithuanian alphabet contains 32 letters. Most of them are Latin, also there are original as *ė*, and borrowed as *š*, *ž* from Check, *ą*, *ę* from Polish [19]. There

are 56 phonemes, consisting of 11 vowels and 45 consonants [1],[16]. Consonants can be soft or hard, for example *b*-b^j, *d*-d^j, except *j* is always soft. Consonants are always soft before *i*, *į*, *y*, *e*, *ę*, *ė*. There are 8 diphthongs that are composed of two vowels *ai*, *au*, *ei*, *ui*, *ou*, *oi*, *ie*, *uo* [11]. Lithuanian has 16 mixed-diphthongs composed of vowel *a*, *e*, *i*, *u*, followed by sonorant *l*, *m*, *n*, *r*, for example *al*, *am*, *an*, *ar*. The language also has 4 affricates *c*,*t͡ʃ*,*d͡ʒ*,*d͡ʒ*. The correspondence between the orthography and phonemes is provided in Table 1, where the International Phonetic Alphabet (IPA) is used to denote the phonemes.

Table 1. Lithuanian orthographic and phonemic correspondence

Vowels		Consonants			
a	/a/	p	/p/,/p^j/	dz	/dz/,/dz^j/
ą	/ɑ/	b	/b/,/b^j/	č	/t͡ʃ/,/$t͡ʃ^j$/
e	/ɛ/	t	/t/,/t^j/	dž	/d͡ʒ/,/$d͡ʒ^j$/
ę	/æ/	d	/d/,/d^j/	m	/m/,/m^j/
ė	/e:/	k	/k/,/k^j/	n	/n/,/n^j/
i	/i/	g	/g/,/g^j/	l	/ɬ/,/l^j/
į, y	/i:/	v	/v/,/v^j/	r	/r/,/r^j/
o	/o:/,/ɔ/	s	/s/,/s^j/	j	/j/
u	/ʊ/	z	/z/,/z^j/	f	/f/,/f^j/
ų, ū	/u:/	š	/ʃ/,/$ʃ^j$/	ch	/x/,/x^j/
		ž	/ʒ/,/$ʒ^j$/	h	/ɣ/,/$ɣ^j$/
		c	/ts/,/ts^j/		

Since Lithuanian has a strong dependency between orthography and pronunciation, grapheme to phoneme rules are implemented easily. In this work our conversion rules are inspired by [16], [7].

3 Experimental setup

3.1 Data set

For the experiments we use data provided by the IARPA Babel program [9], IARPA-babel304b-v1.0b dataset. The data is comprised of spontaneous conversational telephone speech. Two conditions are defined: 1) Full Language Pack (FLP) with about 40 hours of transcribed speech for training, 2) Very Limited Language Pack (VLLP) is a subset of FLP comprising about 3 hours of transcribed data, the remaining data FLP is untranscribed and can be used only in an semi-unsupervised manner [13], [20]. In semi-unsupervised training the recognizer is built with the transcribed portion of the data. Then recognizer is used to generate transcripts for the untranscribed training data. An additional 40 hours of untranscribed data is also available for semi-supervised training. According to the Babel evaluation condition, for the FLP systems only the manual transcriptions of audio training data were used for the language modeling, where as in

case of VLLP, the Web text corpora could also be used for training the language models. In both cases an available untranscribed data could be used for acoustic modeling.

Results are reported on the 10 hour development data set. For keywords spotting (KWS) experiments we use the official 2015 development list provided by NIST. The list has 4079 keywords. There are 412 keywords that not appear in the training data and are considered to be out-of-vocabulary for the FLP condition.

3.2 Baseline recognition systems

In our experiments speech-to-text (STT) systems are built via flat start, when the segmentation is performed without a priori information. The acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities [5]. The models are triphone-based and word position-dependent. Our systems use features provided by BUT [8].

The language model is trained with LIMSI STK toolkit [15]. The models are built using manual transcriptions and additional texts from Web. A single-pass decoding is used. Word lattice is generated by 3-gram LM, and final hypotheses are speech non-speech separation made use of the BLSTM approach described in [6] obtained with consensus decoding. The keywords spotting system is based on the consensus network, and combined word and 7-gram sub-word units are used. The keywords spotting system is described in [10].

3.3 Performance metrics

The performance of speech recognition system is measured using word error rate (WER). Actual term-weighted value (ATWV) is used for the performance of KWS [2]. The keyword specific ATWV for the keyword k at threshold t is computed:

$$ATWV(k,t) = 1 - P_{FR}(k,t) - \beta P_{FA}(k,t) \quad (1)$$

where P_{FR} is the probability of false reject and P_{PFA} of false accept. The constant β is set to 999.9, it mediates the trade off between the false accepts and the false rejects. MTWV is the maximum value computed over all possible thresholds t. The words that are not in the vocabulary are out-of-vocabulary words (OOV), and the rest are in-vocabulary words (INV).

4 Experimental results

We evaluated different phoneme and grapheme systems. STT and KWS results for some FLP grapheme and phoneme systems are showed in Table 3. The WER range for FLP systems was 44.36%-44.81%, and 52.00%-52.64% for VLLP. ATWV and MTWV are reported for combined full-word and 7-gram

sub-word keyword hits. The column Homoph defines the number of homographs and the number of homophones for graphemes and phonemes, respectively. The different grapheme and phoneme sets are described in Table 2. Three additional units are used to model hesitation and silence models. We do not include soft consonants, as it is typically captured by context-dependent triphones. The rare non Lithuanian characters are mapped x→ks, q→k, w→v. For the baseline of phoneme sets, affricates are split into as sequence of two phonemes: c→ts, č→ʧ, dz→dz, dž→dʒ.

Table 2. Grapheme and phoneme systems

System	#Units	Modification from baseline
FLP graph-baseline	35	graphs
FLP phone-baseline	32	phones
FLP phone	36	c→c, č→č, dž→D, dz→Z
FLP phone	38	diph, ou→o,u, oi→o,i
VLLP graph-baseline	33	graphs, c→ts, f→v
VLLP graph	29	z→s, ch→h, ę→e, į,y→y, ū,ų→u:
VLLP phone-baseline	31	phones, f→v
VLLP phone	27	z→s, ch→h, ę→e, į,y→i:, ū,ų→u:

The linguistics argue that affricates and diphthongs can be successfully modeled as separate phonemes. Table 3 shows that merging affricates the slight improvements of ATWV, MTWV are gained with the phoneme based system (phone-36). When diphthongs are modeled as separate units (we split the rare *ou*, *oi* into vowels), system leads to an absolute increase in WER of 0.33% (44.69% vs. 44.36%) Note that the best phonemic system for STT is not the best for KWS. While the best WER result is obtained for phonemic system with units for diphthongs, the best ATWV, MTWV are obtained when affricates are modeled. Phoneme based system gives a slightly lower WER 0.20% absolute compared to the grapheme based system (44.56% vs. 44.36%).

Table 3. WER, ATWV results for graphemic and phonemic FLP systems. ATWV, MTWV computed for combined word and 7-gram sub-word units

System	#Units	Homoph	%WER	ATWV(all/INV/OOV)	MTWV(all/INV/OOV)
graph	35	522	44.56	0.5775/0.5917/0.4497	0.5786/0.5924/0.4716
phone	32	719	44.69	0.5744/0.5897/0.4370	0.5763/0.5912/0.4762
phone	36	718	44.59	**0.5781/0.5923/0.4510**	**0.5801/0.5928/0.4869**
phone	38	718	**44.36**	0.5733/0.5892/0.4309	0.5763/0.5912/0.4600

Table 4 presents WER results for VLLP systems. We investigated the mappings for the rare units as they are poorly modeled. For the baseline systems, only the two rarest units *c*, *f* are mapped (Table 2). Furthermore, we reduce the

number of three more units by mapping *z*, *ch*, *ę* because they are rarely seen in the data. The *į*, *y* and *ū*, *ų* are also mapped as these units define the same sounds but have different representations due to grammar exceptions. It can be observed, that both the grapheme and phoneme based systems perform slightly better when the number of units is reduced. Comparing the best grapheme system with the best phoneme system, the phoneme system obtains an absolute WER gain of 0.20% (52.20% vs. 52.00%). Since the mapping has increased from 1583 to 2418, the limited gain may be due to the increased lexical confusability.

Table 4. WER of VLLP systems

System	#Units	Homoph	%WER
graph	33	493	52.57
graph	29	1583	**52.20**
phone	31	1336	52.27
phone	27	2418	**52.00**

In the FLP experiments we used only the manual manual transcriptions for language modeling. To build VLLP systems, we used Webdata for training language models, and additional untranscribed data for semi-supervised training. Via these experiments, we want to assess the impact of the Webdata and semi-supervised training for the FLP and VLLP systems.

Table 5. WER results for different conditions: only manual transcriptions used for LM training, additional Web texts used, additionally SST used for acoustic models

Set	Hours	Acoustic model	Language model	Lexicon	%OOV	%WER
FLP	40	trn	trn	30k	5.87	44.36
FLP	73	trn + SST	trn	30k	5.87	44.76
FLP	40	trn	trn + webtexts	60k	2.11	**42.41**
FLP	73	trn + SST	trn + webtexts	60k	2.11	42.44
VLLP	3	trn	trn	5.7k	16.42	59.32
VLLP	41	trn + SST	trn	5.7k	16.42	58.98
VLLP	3	trn	trn + webtexts	60k	6.36	53.31
VLLP	41	trn + SST	trn + webtexts	60k	6.36	**52.00**

Table 5 summarizes STT experiments when additional Web texts are used for language modeling, and semi-supervised training for acoustic models. We performed lattice based semi-supervised training, because on some tunng data lattice based training gave 0.60% absolute improvement comparing to 1-best based training (51.08% vs. 50.48%). This is consistent with the study in [3] which reported that lattice based approach gives better results compared to the straight forward 1-best approach. In Table 5 it can be observed that Webdata helps to improve FLP system, absolute gain in WER is 1.95% (44.36% vs. 42.41%). How-

ever, when semi-supervised training is used, FLP system performance decreased both with and without web data.

Contrary to FLP, both Webdata and semi-supervised training help to improve the performance of VLLP system significantly. An absolute gain of 6.01% is obtained adding Webdata for 3 hours system (59.32% vs. 53.31%). Moreover, the improvement of 7.32% is achieved comparing 3 hours system without Webdata and the system with Webdata and semi-supervised training (52.00% vs. 59.32%).

5 Summary

We developed a conversational telephone speech recognition system for the low-resourced language, Lithuanian. We first analyzed the phonemic inventory to identify if phoneme based system outperforms grapheme based system. Using phonemes is found to give only a slight improvement for both FLP and VLLP systems. It is because the Lithuanian language has quite strong dependency between orthographic transcription and phonetic one.

Moreover, we explored the impact of using additional Web texts for training language models, and additional untranscribed data for semi-supervised training. Adding Web texts to FLP system gave an improvement of almost 2%. The VLLP system was improved over 7% absolute using both Webdata and semi-supervised training.

6 Acknowledgments

We would like to thank our Babelon partners for sharing resources (BUT for the bottle-neck features and BBN for the web data), and Grégory Gelly for providing the VADs.

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

1. Ambrazas, V., Garšva, K., Girdenis, A.: Dabartinės lietuvių kalbos gramatika (A Grammar of Modern Lithuanian), Vilnius: MELI (2006)
2. Fiscus, J. G., Ajot, J., Garofolo, J. S., and Doddington, G.: Results of the 2006 spoken term detection evaluation. In Proceedings of ACM SIGIR, 7, 51–55 (2007)

3. Fraga-Silva, T., Gauvain, J.L., and Lamel, L.: Lattice-based Unsupervised Acoustic Model Training. In ICASSP'11, 36th International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 4656-4659, (2011)
4. Gales, M. J. F., Knill, K. M., Ragni, A.: Unicode-based graphemic systems for limited resource languages. ICASSP 2015, (2015)
5. Gauvain, J.L., Lamel L. and Adda G.: The LIMSI broad-cast news transcription system, Speech Communication, 37(1), 89–108 (2002)
6. Gelly, G. and Gauvain J.L.: Minimum Word Error Training of RNN-based Voice Activity Detection. Interspeech 2015, Dresden (2015)
7. Girdenis, A.: Teoriniai lietuvių fonologijos pagrindai (Theoretical Foundations of Lithuanian Phonology), 2nd Edition, Vilnius: Mokslo ir enciklopedijų leidybos inst., (2003)
8. Grézl, F., and Karafiát, M.: Semi-supervised bootstrapping approach for neural network feature extractor training. Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 470-475, (2013)
9. Harper, M.: "IARPA Babel Program," http://www.iarpa.gov/index.php/research-programs/babel
10. Hartmann, W., Le, V. B., Messaoudi, A., Lamel, L., Gauvain, J. L.: Comparing decoding strategies for subword-based keyword spotting in low-resourced languages. Proceedings of Interspeech, Singapore, 2764-2768 (2014)
11. Kazlauskienė, A., Raškinis, G.: Bendrinės lietuvių kalbos garsų dažnumas (The frequency of generic Lithuanian sounds). Respectus Philologicus, 16(21) (2009)
12. Kanthak, S., Ney, H.: Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In ICASSP, 2, 845-848 (2002)
13. Kemp, T., Waibel, A.: Unsupervised training of a speech recognizer: recent experiments. ESCA Eurospeech, 2725–2728 (1999)
14. Killer, M.: Grapheme-based speech recognition. M.S. thesis, Carnegie Mellon University (2003)
15. Lamel, L., Courcinous, S., Despres, J., Gauvain, J.L., Josse, Y., Kilgour, K., Kraft, F., Le, V.B., Ney, H., Nubaum-Thom, M., Oparin, I., Schlüter, R., Schultz, T., Fraga Da Silva, T., Stüker, S., Sundermeyer, M., Vieru, B., Vu. N.T., Waibel, A., and Woehrling, C.: Speech Recognition for Machine Translation in Quaero. IWSLT 2011, 121–128 (2011)
16. Pakerys, A.: Lietuvių bendrinės kalbos fonetika. Enciklopedija, Valiulio Leidykla (The phonetics of generic Lithuanian language. Encyclopedia) (2003)
17. Prasad, R., Matsoukas, S., Kao,C.-L.,Ma, J.Z., Xu,D.-X., Colthurst, T., Kimball, O., Schwartz, R., Gauvain, J.L., Lamel, L., Schwenk, H., Adda, G., and Lefevre., G.: The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System. In InterSpeech, 1645-1648 (2005)
18. Šilingas, D., Laurinčiukaitė, S., Telksnys, L.: Towards Acoustic Modeling of Lithuanian Speech. Proceedings of International Conference SPECOM 2004, 326-333 (2004)
19. Vaišnienė, D., Zabarskaitė, J.: Lithuanuan Language in the Digital Age. Springer, White Paper Series (2012)
20. Zavaliagkos, G., and Colthurst, T.: Utilizing Untranscribed Training Data to Improve Performance. Proc. DARPA Broadcast News Transcription and Understanding Workshop, 301-305 (1998)