Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 107 – 113

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Lithuanian Broadcast Speech Transcription using Semi-supervised Acoustic Model Training

Rasa Lileikytė[a,*], Arseniy Gorin[a], Lori Lamel[a], Jean-Luc Gauvain[a], Thiago Fraga-Silva[b]

[a] *LIMSI, CNRS, Université Paris–Saclay, 508 Campus Universitaire F–91405 Orsay, France*
[b] *Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France*

Abstract

This paper reports on an experimental work to build a speech transcription system for Lithuanian broadcast data, relying on unsupervised and semi-supervised training methods as well as on other low-knowledge methods to compensate for missing resources. Unsupervised acoustic model training is investigated using 360 hours of untranscribed speech data. A graphemic pronunciation approach is used to simplify the pronunciation model generation and therefore ease the language model adaptation for the system users. Discriminative training on top of semi-supervised training is also investigated, as well as various types of acoustic features and their combinations. Experimental results are provided for each of our development steps as well as contrastive results comparing various options. Using the best system configuration a word error rate of 18.3% is obtained on a set of development data from the Quaero program.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Automatic speech recognition; Low-resourced languages; Semi-supervised training; Neural networks; Lithuanian language

1. Introduction

With only about 3.5 million speakers, Lithuanian is one of the least spoken languages in Europe. It belongs to the Baltic subgroup of Indo-European languages. Lithuanian writing is based on the Latin alphabet, with some accentuated characters. It has a complex stress system and flexible word order[1]. The language is highly inflected. All these factors result in a large dictionary, a high out-of-vocabulary rate and the lack of data for language modeling[2].

Few studies report on speech recognition for the Lithuanian language, in part due to the sparsity of linguistic resources. Lithuanian systems for isolated word recognition are described in[3,4]. Studies addressing

* Corresponding author. Tel.: +33-(0)1-69-85-81-82.
E-mail address: lileikyte@limsi.fr, rasalileikyte@gmail.com

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.037

conversational telephone speech recognition for the Lithuanian language are reported in[5,6]. To our knowledge, there are just a few works addressing broadcast speech in Lithuanian. In[7,8] the systems for broadcast data were trained on only 9 hours of transcribed data. The broadcast news transcription system developed in the context of the Quaero program by LIMSI and Vocapia Research[2] was trained without any manually annotated data. This paper reports on extensions of this initial work.

In the next section the data set used for the experiments is presented, followed by a description of the baseline system developed in the Quaero program. The revised training process and five techniques aiming to improve the acoustic models are then described along with a presentation of the results obtained.

2. Data set

All experiments use the data collected during the Quaero program[9]. This corpus contains about 440 hours of raw audio data[10]. It is comprised of Lithuanian broadcast news speech, downloaded from the following channels: *Žinių radijas*, *Lietuvos nacionalinis radijas ir televizija*, and *Lietuvos radijas* (www.ziniuradijas.lt, www.lrt.lt, www.radijas.fm). The audio data were automatically partitioned[11], resulting in about 360 hours of audio segments detected as containing speech. The text corpus used to train the language model includes about 9 millions words of texts collected from the Web, with a focus on sources containing transcripts of broadcast news and interviews such as *15min*, and *Aukštaitijos internetinė naujienų agentūra* (www.15min.lt, www.aina.lt). The text normalization follows the process described in[12,13]. A development data set of 3 hours with manual transcriptions is used to evaluate and compare the models. A second data set of 3 hours with manual transcriptions is used for the final evaluation. Finally a third 3 hour data set with manual transcripts, which was never used in the initial Quaero work[2], serves here as training data to improve the bootstrapping of the semi-supervised training process.

3. Baseline system and results

Our baseline automatic speech recognition system refers to the system developed during the Quaero program. It uses left-to-right 3-state hidden Markov models (HMMs) with Gaussian mixture observation densities, in total about 10k tied states with about 32 components per state[10]. The triphone-based phone models are word position-dependent, gender-dependent and speaker-adaptive trained (SAT). The system is bootstrapped with context-independent English seed phone models trained on a large amount data. This language transfer is obtained by mapping the Lithuanian phonemes to a close English counterpart[14]. Then, unsupervised training is performed using 360 hours of untranscribed data[2,11,15]. The features are extracted from a bottleneck layer of multilayer perceptron (MLP) with 3 hidden layers trained on Russian broadcast data[16,17] using TRAP-DCT acoustic features. The bottleneck (BN) features are augmented with perceptual linear prediction (PLP) and pitch features.

The Lithuanian alphabet contains 32 Latin based letters, with 12 vowels and 20 consonants. The system uses a 25 phone set, containing 6 vowels, 16 consonants and 3 special phones. The long and short vowels are merged, and affricates are split into a sequence of two phonemes. The non Lithuanian characters appearing in the corpus are mapped to Lithuanian ones, e.g. x→ks, q→k, w→v. A 200k word list was created by selecting the most frequent words in the text corpus. The out-of-vocabulary (OOV) rate on the development data is 3.4%. Given the close correspondence between the orthographic and phonemic realization in Lithuanian a set of grapheme-to-phoneme conversion rules was used to generate the pronunciations of the words in the 200K lexicon[10].

Four-gram back-off language models (LM) with Kneser-Ney smoothing were trained on the text corpus described in Section 2. LMs were built for each source of the training texts, and then interpolated using the EM algorithm to minimize the perplexity of the development set. For each speech segment a word lattice is generated, the final hypotheses are then obtained using consensus decoding[18]. The results on the development data obtained for the initial work are shown in Table 1.

As described in[19] an iterative procedure was used for unsupervised training, roughly doubling the amount of raw audio data in each iteration. Stage A in Table 1 is the result after the 4th unsupervised training

Table 1. Baseline system development: word error rate (WER) at the different development stages.

Stage	Features	#Hours (untrans)	%WER
A	PLP+F0	91	50.8
B	PLP+F0	290	39.6
C	PLP+F0+BN(TRAP)	290	33.9
D	PLP+F0+BN(TRAP)	368	28.0

iteration, where the acoustic models (PLP-based HMM-GMM) were trained on about 90 hours of automatically transcribed speech data. The WER with these models is about 50%. In the next iteration (stage B), the amount of training data was roughly tripled, leading to a 10% absolute (22% relative) reduction in WER. Using the same automatic transcripts with the TRAP bottleneck MLP features (stage C) gives almost an additional 14% relative WER reduction. Re-transcribing all of the data with the BN models increases the amount of automatically transcribed speech and gives a 17% relative reduction in WER (28%).

4. Revised training process

With the goal of significantly improving the baseline results reported in Section 3, we completely revised the system development recipe while keeping the language model unchanged. The following changes were introduced in the acoustic model training:

- use the left out 3h data set with transcriptions to provide a better bootstrap for the semi-supervised training process;
- replace the phonemic lexicon by a graphemic lexicon to simplify system development;
- use a deep neural network (DNN) to estimate the HMM state likelihoods replacing the GMMs;
- increase the weight of this small transcribed data set to further improve the model accuracy;
- take advantage of the manually transcribed data to discriminatively train the acoustic models;
- combine various acoustic feature sets (PLP, TRAP and filterbank bottleneck features).

Table 2 gives the progression of results as well as some contrastive results obtained while revising the acoustic model training procedure.

4.1. Bootstrapping with 3h of transcribed data

A small amount (3 hours), of previously unused, transcribed data from the Quaero project served to construct bootstrap models, i.e., in contrast to the baseline system, there is no need for language transfer. The GMM-HMM systems are built via a flat start training. The HMM architecture is similar to the one described in Section 3[20]. For all experiments, the triphone-based models are word position-dependent, like in the baseline. The system uses PLP and F0 features (later denoted as PLP). Audio segmentation and speaker diarization are carried out using the audio partitioner as described in[11].

For semi-supervised training (SST) the same corpus of untranscribed broadcast audio is used, as described in Section 2. The initial GMM-HMM trained on 3 hours of manually transcribed data is used to decode the untranscribed data set. SST is carried out in an incremental manner[21,22]. First, an 11 hour subset of the untranscribed audio is automatically transcribed. The hypothesized transcriptions are used as ground truth reference transcripts for re-training the models. Then at each iteration the amount of untranscribed data is increased and the data from previous iterations is re-decoded with the new models.

After semi-supervised training is finished, the final GMM-HMM model is trained using 9 frame spliced features, followed by linear discriminant analysis (LDA) for dimensionality reduction (40 dimension) and speaker adaptive training (SAT). As shown in part B of Table 2, the WER of the initial phonemic based system is 36.1%.

Table 2. Word error rate (WER) of various systems described in the work.

#	Model	Features	#Hours trans	#Hours + untrans	SST	graph	% WER SI	% WER SAT
A	GMM (baseline)	PLP + BN TRAP	0	+ 360	no	no	–	28.0
B	GMM	PLP	3		no	no	36.1	–
	GMM	PLP	3		no	yes	36.8	–
	GMM	PLP	3	+ 360	1^{st} pass	no	30.5	26.6
	GMM	PLP	3	+ 360	1^{st} pass	yes	31.0	26.3
C	DNN CE	PLP	3	+ 360	1^{st} pass	no	–	22.1
	DNN CE	PLP	3	+ 360	1^{st} pass	yes	–	22.0
	DNN sMBR	PLP	3	+ 360	1^{st} pass	yes	–	21.5
D	DNN CE	PLP	3x10	+ 360	1^{st} pass	yes	–	21.7
	DNN sMBR	PLP	3x10	+ 360	1^{st} pass	yes	–	20.6
E	GMM	BN PLP	3		1^{st} pass	yes	25.8	–
	GMM	BN PLP	3	+ 360	2^{nd} pass	yes	23.6	21.6
F	DNN CE	PLP	3x10	+ 360	2^{nd} pass	yes	–	19.7
	DNN sMBR	PLP	3x10	+ 360	2^{nd} pass	yes	–	19.0
G	DNN sMBR	BN PLP (46)	3x10	+ 360	2^{nd} pass	yes	19.6	19.2
	DNN sMBR	BN TRAP (42)	3x10	+ 360	2^{nd} pass	yes	19.3	18.8
	DNN sMBR	BN FBANK (46)	3x10	+ 360	2^{nd} pass	yes	19.4	**18.6**
	DNN sMBR	BN PLP + BN TRAP	3x10	+ 360	2^{nd} pass	yes	19.2	18.6
	DNN sMBR	BN PLP + BN TRAP + BN FBANK	3x10	+ 360	2^{nd} pass	yes	19.2	**18.3**

4.2. Graphemic lexicon

For the Lithuanian language there is a quite strong dependency between the orthographic transcription and the phonetic form, which simplifies the creation of pronunciation dictionaries. In this work we experiment with a pronunciation dictionary based on graphemes, as[5] showed that for a conversational telephone speech task phonemes only gave a slight improvement compared to graphemes. Each orthographic character is modeled as a separate grapheme, for at total of 34 units, including two pseudo phone units representing silence and fillers.

The first two entries in Table 2 part B compare the phoneme based lexicon from the baseline system and the grapheme based lexicon using an GMM-HMM trained on the initial 3 hour set of data. This comparison shows that the difference in WER is less than 2% relative (WER of 36.1% vs 36.8%). The WER of the grapheme-based system is reduced by 5.8% absolute after SST and by additional 4.7% after applying LDA+SAT. Last two lines of Table 2 part B compare grapheme and phoneme based lexicons when using SST and LDA+SAT. Again, the relative difference is less than 2% WER, with a slightly better result achieved using graphemic lexicon.

4.3. DNN

In this step of the revised training procedure, a DNN is used to estimate the HMM state likelihoods replacing the GMMs[23]. Data weighting is also explored to compensate the lack of transcribed training data, followed by discriminative training of the DNN.

Training makes use of the alignments produced by the SAT GMM-HMM system described earlier. A 6-layer DNN with 10M parameters is trained with the cross-entropy (CE) criterion[24], using the same features used for GMM-HMM system: 9 frame splicing and LDA+SAT. Sequence-discriminative training is applied using the state-level minimum Bayes risk (sMBR) criterion[25]. Part C of Table 2 shows that an absolute improvement of 0.5% is obtained using sMBR vs CE training. This improvement is much smaller than that observed with supervised training.

There is a large difference between the amount of manually transcribed and automatically transcribed data (3 versus 360 hours) in our training corpus. The transcripts produced in semi-supervised manner are

naturally erroneous, which can lead to degradation of the system performance. To address this concern, following the positive results reported in[26], the impact of weighting the supervised/unsupervised data sets for DNN training was explored. In addition, we also use only the manually transcribed data in the sMBR fine-tuning stage. Table 3 shows that the largest improvement is obtained by using 10 copies of manually transcribed data. In this case the WER is reduced by 0.9% absolute compared to the sMBR trained DNN (21.5% vs 20.6%).

Table 3. WER for DNN-HMM using the untranscribed and from 2 to 30 copies of the manually transcribed data. Only the manually transcribed data are used for sMBR.

# of copies	2x	5x	10x	15x	20x	30x
DNN	21.9	21.8	21.7	22.0	22.2	22.2
+sMBR	21.1	20.7	**20.6**	20.7	21.0	21.2

4.4. Refining the transcriptions

In this section bottleneck (BN) features are explored to further improve the accuracy, and subsequently serve to refine the automatic transcriptions used for semi-supervised training. The training procedure is repeated, replacing the PLP features with bottleneck features.

A DNN with a bottleneck layer (DNN BN) was trained on the same data. The architecture of the network is equivalent to the DNN described in Section 4.3, except the next-to-last hidden layer is replaced with a small layer with only 46 units.

After extracting BN features we repeat the SST process described in Section 4.1, but using BN features instead of PLP. After training an initial model on the manually annotated data, we refine the automatic transcriptions used for semi-supervised training. Part E of Table 2 summarizes the performance of the GMM-HMM trained with 3h of manually annotated data using BN features, and the GMM-HMM performance after repeating SST.

Comparing the PLP based and BN based GMM-HMM systems (parts B and E of Table 2), the latter obtains an absolute WER improvement of 11.0% (36.8% vs 25.8%) when trained on the 3h data set. SST reduces the WER by 7.4% absolute (31.0% vs 23.6%).

The refined semi-supervised transcripts produced with the BN features were used to train a DNN-HMM. Only the hypothesized transcripts were changed, the previous alignments and PLP features were used for DNN training. Comparing parts D and F of Table 2, it can be observed that new transcripts lead to absolute WER reduction of 1.6% (20.6% vs to 19.0%) when the same alignments and training conditions are used for DNN-HMM trained with sMBR.

4.5. Feature set combination

In this section we investigate BN features adapted using constrained maximum likelihood linear regression (CMLLR) as an input to a hybrid DNN. In addition to PLP features, TRAP and filterbank (FBANK) features are explored for BN DNN training, as well as the combination of several BN features.

As in the previous section, all systems use 10 copies of manually transcribed data set and the automatically transcribed data. The same DNN architecture and the same alignments are used in order to be able to compare the improvements coming from replacing PLP features with BN features. BN DNNs are trained with cross-entropy criterion and then fine-tuned with sMBR using only the manually transcribed data.

Three types of acoustic features were compared: PLP features used earlier; TRAP features extracted using 19-band Mel spectrogram with 30 ms window and 10 ms offset, giving 168 features after applying discrete cosine transform[17]; and FBANK features composed of 24 Mel filterbanks, augmented with c0 and pitch. All features are mean and variance normalized. In addition, for the PLP and FBANK features 13 frames are spliced. The bottleneck dimensions are 46 for the PLP and FBANK features, and 42 for the TRAP features.

For the PLP+TRAP+FBANK combination the resulting vector containing 134 values is used as the input to the hybrid DNN.

Part G of Table 2 gives the WER achieved with hybrid DNN systems using various BN features and their combinations. The system with PLP BN features obtains a WER of 19.2%, having roughly the same performance to the hybrid system trained using CMLLR adapted PLP features (19.0% WER, last entry in part F). The DNN trained with FBANK BN features obtains the best WER of 18.6% with a single feature set. A disappointingly small improvement is achieved by combining 3 sets of BN features (PLP+TRAP+FBANK) and adapted with CMLLR, in contrast to the large gain achieved with feature combination for the baseline system. Overall this system obtains a relative WER reduction of of 34.6% (18.3% vs 28.0%) compared to the baseline. This final model was used to decode the Quaero evaluation data set, and achieved a WER of 19.3% WER, which is 28% relative improvment over the baseline which had a WER of 26.9%.

5. Summary

This paper has reported on developing a speech-to-text transcription system for the low-resourced Lithuanian language. The goal was to improve over the baseline HMM-GMM system developed in the Quaero program. To do so the acoustic model training procedure was revised to bootstrap the semi-supervised training process with a flat start (rather than cross-language transer) using a small amount of transcribed data. The phonemic lexicon was replaced by a graphemic one to simplify system development. A DNN was used to estimate the HMM state likelihoods replacing the GMMs, and data weighting was used to increase the relative importance of the small transcribed data set enabling discriminative training of the acoustic models. The final system achieved a WER of 19.3% on the Quaero evaluation data, which is about 28% relative than the Quaero system which served as a baseline for this work. The small gain achieved with feature combination were disappointing compared to the improvement observed with the baseline HMM-GMM system.

References

1. Vaišnienė D, Zabarskaitė J, Rehm G, Uszkoreit H. Lithuanuan language in the digital age, Springer. 2012.
2. Lamel L. Unsupervised acoustic model training with limited linguistic resources. In: *ASRU*; 2013.
3. Lipeika A, Lipeikienė J, Telksnys L. Development of isolated word speech recognition system. *Informatica* 2002. **13**:37–46.
4. Raškinis G, Raškinienė D. Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. *Informatica*; 2003.**14**:75–84.
5. Lileikyte R, Lamel L, Gauvain JL. Conversational telephone speech recognition for Lithuanian. In: *SLSP*; 2015, p. 164–172.
6. Gales MJF, Knill KM, Ragni A. Unicode-based graphemic systems for limited resource languages. In: *ICASSP*; 2015, p. 5186–5190.
7. Laurinčiukaitė S, Lipeika A. Syllable-phoneme based continuous speech recognition. *Elektronika ir Elektrotechnika* 2006;**70**:91–94.
8. Šilingas D, Laurinčiukaitė S, Telksnys L. Towards acoustic modeling of Lithuanian speech. In: *SPECOM*; 2004, p. 326–333.
9. Lamel L. Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In: *Baltic HLT*; 2012. p. 1–8.
10. Report ID.CTC.QPR23 of the Quaero Program. 2013.
11. Gauvain JL, Lamel L, Adda G. Partitioning and transcription of broadcast news data. In: *ICSLP*; 1998. p. 1335–1338.
12. Adda G, Adda-Decker M, Gauvain JL, Lamel L. Text normalization and speech recognition in French. In: *Eurospeech*; 1997. p. 56–59.
13. Adda-Decker M, Adda G, Lamel L. Investigating text normalization and pronunciation variants for German broadcast transcription. In: *Interspeech* ; 2000. p. 266–269.
14. Lamel L, Gauvain JL. Automatic processing of broadcast audio in multiple languages. In: *11th European Signal Processing Conference, 2002*; 2002. p. 1–4.
15. Lamel L, Gauvain JL, Adda G. Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language* 2002;**16**:115–129.
16. Hermansky H, Ellis DW, Sharma S. Tandem connectionist feature extraction for conventional HMM systems. In: *ICASSP*; 2000. p. 1635–1638.
17. Grézl F, Fousek P. Optimizing bottle-neck features for LVCSR. In: *ICASSP*; 2008. p. 4729–4732.
18. Mangu L, Brill E, Stolcke A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language* 2000;**14**:373–400.
19. Lamel L, Vieru B. Development of a speech-to-text transcription system for Finnish. *SLTU* 2010;**10**:62–67.
20. Gauvain JL, Lamel L, Adda G. The LIMSI broadcast news transcription system. *Speech communication* 2002; **37**:89–108.

21. Lamel L, Gauvain JL, Adda G. Unsupervised acoustic model training. In: *ICASSP*; 2002. p. 877-880.
22. Fraga-Silva T, Gauvain JL, Lamel L. Lattice-based unsupervised acoustic model training. In: *ICASSP*; 2011. p. 4656–4659.
23. Dahl GE, Yu D, Deng L, Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *Audio, Speech, and Language Processing* 2012;**20**:30–42.
24. Zhang X, Trmal J, Povey D, Khudanpur S. Improving deep neural network acoustic models using generalized maxout networks. In: *ICASSP*; 2014. p. 215–219.
25. Vesely K, Ghoshal A, Burget L, Povey D. Sequence-discriminative training of deep neural networks. In: *Interspeech*; 2013. p. 2345–2349.
26. Vesely K, Hannemann M, Burget L. Semi-supervised training of deep neural networks. In: *ASRU*; 2013. p. 267–272.