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