Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Conversational telephone speech recognition for Lithuanian☆

Q2 Rasa Lileikytė*, Lori Lamel, Jean-Luc Gauvain, Arseniy Gorin
Q3 *LIMSI, CNRS, Université Paris-Saclay, 508 Campus Universitaire, Orsay F-91405, France*

Received 23 March 2016; received in revised form 24 April 2017; accepted 28 November 2017
Available online xxx

Abstract

The research presented in the paper addresses conversational telephone speech recognition and keyword spotting for the Lithuanian language. Lithuanian can be considered a low e-resourced language as little transcribed audio data, and more generally, only limited linguistic resources are available electronically. Part of this research explores the impact of reducing the amount of linguistic knowledge and manual supervision when developing the transcription system. Since designing a pronunciation dictionary requires language-specific expertise, the need for manual supervision was assessed by comparing phonemic and graphemic units for acoustic modeling. Although the Lithuanian language is generally described in the linguistic literature with 56 phonemes, under low-resourced conditions some phonemes may not be sufficiently observed to be modeled. Therefore different phoneme inventories were explored to assess the effects of explicitly modeling diphthongs, affricates and soft consonants. The impact of using Web data for language modeling and additional untranscribed audio data for semi-supervised training was also measured. Out-of-vocabulary (OOV) keywords are a well-known challenge for keyword search. While word-based keyword search is quite effective for in-vocabulary words, OOV keywords are largely undetected. Morpheme-based subword units are compared with character n-gram-based for their capacity to detect OOV keywords. Experimental results are reported for two training conditions defined in the IARPA Babel program: the full language pack and the very limited language pack, for which, respectively, 40 h and 3 h of transcribed training data are available. For both conditions, grapheme-based and phoneme-based models are shown to obtain comparable transcription and keyword spotting results. The use of Web texts for language modeling are shown to significantly improve both speech recognition and keyword spotting performance. Combining full-word and subword units leads to the best keyword spotting results.

Keywords: Conversational telephone speech; Lithuanian; Speech-to-text; Keyword spotting

1. Introduction

Lithuanian belongs to the Baltic subgroup of Indo-European languages and is one of the least spoken European languages, with only about 3.5 million speakers. Although the language was standardized during the late 19th and early 20th centuries, most of the phonetic and morphological features were preserved (Vaišnienė et al., 2012). The language is characterized by a rich inflection, a complex stress system, and a flexible word order. Lithuanian is

☆ This paper has been recommended for acceptance by Roger K. Moore.
* Corresponding author.
E-mail address: lileikyte@limsi.fr (R. Lileikytė).

http://dx.doi.org/10.1016/j.csl.2017.11.005

written using the Latin alphabet with some additional language specific characters, as well as some 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 of Lithuania by 3 million speakers. Samogitian is spoken in the west of the country by only about 0.5 million speakers.

This paper reports on research work aimed at developing conversational telephone speech (CTS) recognition and keyword spotting (KWS) systems for the Lithuanian language. Speech recognition systems making use of statistical acoustic and language models are typically trained on large data sets. Three main resources are needed: (1) telephone speech recordings with corresponding transcriptions for acoustic model training, (2) written texts for language modeling, and (3) a pronunciation dictionary.

There have been only a few studies reporting on speech recognition for Lithuanian, in part due to the sparsity of the available linguistic e-resources. Systems for isolated word recognition are described in Lipeika et al. (2002), Maskeliūnas et al. (2015), Filipovič and Lipeika (2004), Raškinis and Raškinienė (2003), Vaičiūnas and Raškinis (2006). In Laurinčiukaitė and Lipeika (2015), Šilingas et al. (2004), Šinlingas (2005), Lithuanian broadcast speech recognition systems were trained on 9 h of transcribed speech, where in Laurinčiukaitė and Lipeika (2015) syllable sets and in Šilingas et al. (2004), Šinlingas (2005) different phonemic units were investigated. In the context of the Quaero program (www.quaero.org), a transcription system for broadcast audio in Lithuanian was developed without any manually transcribed training data and achieved 28% word error rate (WER) (Lamel, 2013). Using only 3 h of transcribed audio data and semi-supervised training, this result was later improved to 18.3% (Lileikytė et al., 2016). In Gales et al. (2015) a unicode-based graphemic system for the transcription of conversational telephone speech in Lithuanian is described. The system, developed within the IARPA Babel program, obtained a WER of 68.6% with 3 h of transcribed training data, and of 48.3% using 40 h of transcribed training data.

Transcribing conversational telephone speech is a more challenging task than transcribing broadcast news, which is predominantly comprised of prepared speech by professional speakers. In spontaneous speech, speaking rates and styles vary across speakers and grammar rules are not strictly followed. Example phrases illustrating some common phenomena found in casual speech are given in Table 1. Hesitations and filler sounds occur frequently in conversational speech, appearing in 30% of the speaker turns (counted in the training transcripts). Disfluencies and/or unintelligible words are marked in 25% of the speaker turns. Moreover, the audio signal has a reduced bandwidth of 3.4 kHz and can be corrupted by noise and channel distortion.

The research reported in this paper was carried out in the context of IARPA Babel project using the IARPA-babel304b-v1.0b corpus. The data were collected in a wide variety of environments, and have a broad range of speakers. There is a wide distribution of speakers with respect to gender, age and dialect. The audio were recorded in various conditions such as on the street, in a car, restaurant or office, and with different recording devices such as cell phones and hands-free microphones.

This study uses the same training and test resources as (Gales et al., 2015) for two conditions: the full language pack (FLP) with approximately 40 h of transcribed telephone speech and the very limited language pack (VLLP) comprised of only a 3 h subset of the FLP transcribed speech. An additional 40 h set of untranscribed data was available for semi-supervised training. A 26 million word text corpus, collected from the Web (Wikipedia, subtitles and other sources) and filtered by BBN (Zhang et al., 2015) was provided. Although the harvesting process searches the Web for texts containing n-grams that are frequent in the transcribed audio, the recovered texts are for the most part quite different from conversational speech. The available resources for acoustic and language model training are

Table 1
Examples of conversational telephone speech phrases.

Event	Example
Hesitations	*aha* tai tada aš turėčiau eiti *mmm* pas draugę
	aha so then I should go *mmm* to a friend
Filler words	*nu* bet iš ryto *žinai* aštuntą valandą
	yeah but in the morning *you know* at eight o'clock
Word fragments	susitikim *šeštad-* ne sekmadienį
	let's meet on *sat-* no on Sunday
Word repetitions	*taip taip taip* papietaukime *prie prie* parko
	yeah yeah yeah let's have a lunch *near near* a park

Please cite this article as: R. Lileikytė et al., Conversational telephone speech recognition for Lithuanian, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.11.005