Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 49 (2018) 71–82

www.elsevier.com/locate/csl

# Conversational telephone speech recognition for Lithuanian☆

Rasa Lileikytė*, Lori Lamel, Jean-Luc Gauvain, Arseniy Gorin

*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 6 December 2017

**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 units 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 is 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
0885-2308/

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 the IARPA Babel program 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

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 |

quite different from conversational speech. The available resources for acoustic and language model training are very small compared to the 2000 h of transcribed audio and over a billion words of text that are available for the American English conversational telephony task (Prasad et al., 2005).

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 pronunciations are easily derived from the orthographic form. Grapheme-based systems have been shown to work reasonably well for various languages, such as Dutch, German, Italian, and Spanish (Kanthak and Ney, 2002; Killer et al., 2003). Yet, some languages, such as English, have a weak correspondence between graphemes and phonemes, and using graphemes leads to a degradation in system performance. Phoneme-based systems usually provide better results as they better represent the speech production. However, designing the pronunciation dictionary (Adda-Decker and Lamel, 2000) or a set of grapheme-to-phoneme rules requires linguistic expertise, making it a costly process. The Lithuanian language has quite a strong dependency between the orthographic and the phonemic forms making it relatively easy to write grapheme-to-phoneme conversion rules in comparison to the English language that requires numerous exceptions.

This article is the extension of our previous work (Lileikyte et al., 2015). New research results are reported, providing keyword spotting results for both the FLP and VLLP conditions and an analysis of keyword spotting performance with a focus on out-of-vocabulary (OOV) keyword detection. In this study two techniques of enhancing the detection of the OOV keywords are explored: using Web resources to augment the lexicon and language model, and using subword units to enhance KWS. We investigate two types of subword units: character N-grams and morpheme subwords. The use of full-word and subword units for KWS is compared. Moreover, we explore the impact of acoustic model data augmentation using semi-supervised training. To see the benefits of using augmented texts for keyword spotting, we analyze if OOV words become in-vocabulary (INV) or remain OOV.

This study addresses the following questions: (1) which set of phonemic units should be used? (2) is a phoneme-based system better than grapheme-based one? (3) how much improvement can be obtained by using untranscribed audio and Web texts for model training? (4) how much do subword units improve keyword spotting?

The next section describes the phonemic inventory of the Lithuanian language. Section 3 describes the experimental conditions. An overview of the speech-to-text and keyword spotting systems is provided in Section 4. Experimental results comparing different sets of phonemes and graphemes are given in Section 5, and Section 6 investigates semi-supervised training and the use of Web texts for speech recognition. Section 7 focuses on the improvement of out-of-vocabulary keyword detection. Finally, in Section 8 this work is summarized and some conclusions are drawn.

## 2. Lithuanian phonemic inventory

The Lithuanian alphabet contains 32 letters. While most of them are Latin, there is also *ė*, and some borrowed letters from Czech (*š, ž*), and Polish (*ą, ę*). Lithuanian is generally described as having 56 phonemes, comprised of 11 vowels and 45 consonants (Pakerys, 2003). Consonants are classified as soft (palatalized) or hard (not palatalized), where the soft consonants always occur before certain vowels (*i, į, y, e, ę, ė*). There are 8 diphthongs that are composed of two vowels (*ai, au, ei, ui, ou, oi, ie, uo*) and 16 mixed diphthongs composed of a vowel (*a, e, i, u*), followed by a sonorant (*l, m, n, r*), for example, *al, am, an, ar* (Kazlauskienė and Raškinis, 2009). There are also 4 affricates (*c, č, dz, dž*). The correspondence between the orthography and phonemes is provided in Table 2, where the International Phonetic Alphabet (IPA) is used to denote the phonemes. The grapheme-to-phoneme conversion rules used in this work were inspired by Pakerys (2003), Girdenis (2003). In Lithuanian, lexical tone can have a distinctive function and differentiate words (Girdenis, 2003). Long syllables can have either falling (acute) or rising (circumflex) tone, and short syllables can be stressed or not. Lexical tone distinctions were not taken into account in our studies. In our implementation there are 43 grapheme-to-phoneme rules of which 11 are contextual, as well as contextual rules for consonant palatalization. These concern diphthongs, affricates, long/short vowel distinctions and assimilation. Three example words with the phonemic pronunciations are shown in Fig. 1. In the first example the initial ‘l’ is palatalized since it precedes an ‘i’. The four consonants in the second example are also palatalized, and the ‘b’ is pronounced as a /p/ due to assimilation with the following voiceless stop. The voicing assimilation of the ‘k’ in the third example goes from voiceless to voiced.

Table 2
Lithuanian orthographic and phonemic correspondence.

| Vowels | | Consonants | | | |
|---|---|---|---|---|---|
| a | /a/ | p | /p/, /pʲ/ | dz | /dz/, /dzʲ/ |
| ą | /ɑ/ | b | /b/, /bʲ/ | č | /ʧ/, /ʧʲ/ |
| e | /ɛ/ | t | /t/, /tʲ/ | dž | /dʒ/, /dʒʲ/ |
| ę | /æ/ | d | /d/, /dʲ/ | m | /m/, /mʲ/ |
| ė | /e:/ | k | /k/, /kʲ/ | n | /n/, /nʲ/ |
| i | /i/ | g | /g/,/gʲ/ | l | /ɬ/, /lʲ/ |
| į, y | /i:/ | v | /v/, /vʲ/ | r | /r/, /rʲ/ |
| o | /o:/,/ɔ/ | s | /s/, /sʲ/ | j | /j/ |
| u | /ʊ/ | z | /z/, /zʲ/ | f | /f/, /fʲ/ |
| ų, ū | /u:/ | š | /ʃ/,/ʃʲ/ | ch | /x/, /xʲ/ |
| | | ž | /ʒ/, /ʒʲ/ | h | /ɣ/, /ɣʲ/ |
| | | c | /ts/, /tsʲ/ | | |

| | |
|---|---|
| l̲i̲etus (rain) | $l^{j}$iɛtʊs |
| **d**i̲**r**b**t**i̲ (to work) | **d**$^{j}$i**r**$^{j}$**p**$^{j}$**t**$^{j}$i |
| su**k**d̲amas (rotating) | sʊ**g**damas |

Fig. 1. Example of contextual grapheme-to-phoneme (g2p) rules.

## 3. Corpus and task description

This section describes the speech and text corpora, and the experimental conditions used in this work. All the experiments reported in this paper use data provided by the IARPA Babel program Harper, more specifically the IARPA-babel304b-v1.0b data set. The data are comprised of spontaneous telephone conversations, and as mentioned earlier, two conditions are considered:[1] (1) the full language pack (FLP) with about 40 h of transcribed speech for training, and 2) the very limited language pack (VLLP), which is a 3-h subset of the FLP data. For the VLLP condition the remaining 37 h of FLP data are considered as additional untranscribed data that can be used only in a semi-supervised manner (Lamel et al., 2002; Kemp and Waibel, 1999; Zavaliagkos and Colthurst, 1998). The data subsets are shown in Fig. 2. For both conditions an additional set of 40 h of untranscribed data is also available for semi-supervised training.

According to the Babel 2015 evaluation conditions, the FLP language model training was restricted to the transcriptions of audio training data, whereas Web texts could be included in the language model training for the VLLP systems. In Sections 6 and 7 experiments are reported assessing the importance of using the Web data for language modeling for both the VLLP and FLP conditions.

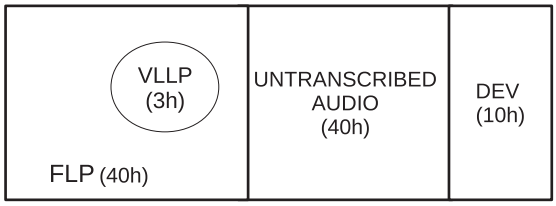


Fig. 2. Available data for FLP and VLLP conditions.

[1] The IARPA Babel program started in 2012, with the first keyword spotting evaluation held in the spring of 2013 (http://www.nist.gov/itl/iad/mig/openkws13.cfm). The VLLP condition was introduced in the third phase of the program and was one of the primary tasks in the 2015 evaluation (http://www.nist.gov/itl/iad/mig/openkws15.cfm).

All results are reported on the official Babel 10 h development data set. For the keyword spotting (KWS) experiments, the official 2015 year list of development keywords provided by the Babelon and LORELEI teams was used (Cui et al., 2014). The development keyword list contains 4079 keywords, where a keyword may be a single word or a sequence of words. If any word in the keyword list is out-of-vocabulary (OOV) then the keyword is considered OOV with respect to the system's vocabulary. The in-vocabulary (INV) and OOV keywords are therefore different for the FLP and VLLP conditions. Based on the vocabulary of their respective transcriptions, there are 412 OOV keywords for the FLP condition, and 474 for the VLLP one.

The speech recognition performance is measured with the commonly used word error rate metric. Keyword spotting results are reported in terms of the maximum term-weighted value (MTWV) (Fiscus et al., 2007). In the Babel program the actual term-weighted value (ATWV) is also used. The keyword specific ATWV for the keyword $k$ at a threshold $t$ is defined as follows:

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

where $P_{FR}$ is the probability of a false reject and $P_{FA}$ of a false accept. The coefficient $\beta$ is set to 999.9. This constant controls the trade-off between the false accepts and the false rejects. The decision threshold is the same for all queries. MTWV is the maximum value computed over all possible thresholds $t$.

## 4. Speech-to-text and keyword spotting

The acoustic models for the speech-to-text (STT) systems are built via a flat start training, where the initial segmentation is performed without any a priori information. The acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (Gauvain et al., 2002). The models are triphone-based and word position-dependent. The system uses discriminatively trained stacked bottleneck acoustic features extracted from a deep neural network that were provided by BUT (Grézl and Karafiát, 2014). In order to abide by the evaluation rules, different features were provided for the FLP and VLLP conditions, in the latter case being trained on multilingual data and fine-tuned using the Lithuanian data. Semi-supervised acoustic model (AM) training[2] with 77 h of raw audio is used for the VLLP condition.

The language model (LM) is trained with the LIMSI STK toolkit. For the FLP condition a 3-gram language model is built using only the manual transcriptions. In the case of the VLLP, the 3-gram language model is trained using both the manual transcriptions and the Web texts.

Prior to transcription, speech/non-speech segmentation is carried out using the bidirectional long short-term memory recurrent neural network as described in Gelly and Gauvain (2015). For each speech segment, a word lattice is generated and the final word hypothesis is obtained via consensus decoding (Mangu et al., 2000).

Keyword search is carried out on the consensus network, ignoring word boundaries. The keyword scores are normalized using keyword-specific thresholding and exponential normalization (Karakos et al., 2013). Full-word and character 7-gram cross-word subword units are used, as described in Hartmann et al. (2014).

## 5. Phoneme-based and grapheme-based systems

Several phoneme-based and grapheme-based systems are evaluated, contrasting different sets of elementary units and mappings for rarely seen units. One contrast explores explicitly modeling complex sounds such as affricates and diphthongs as a single unit or splitting them into a sequence of units. Another compares explicitly modeling the soft consonants as opposed to simply allowing them to be contextual variants of their hard counterparts.[3]

The phoneme and grapheme sets studied in this work are listed in Table 3. Three special symbols are used to represent breath noise, hesitations and silence.

[2] In semi-supervised training a speech recognizer is used to generate transcripts for the untranscribed training data.

[3] The dictionary provided by Appen as part of the Babel language pack explicitly represents stress and soft consonants. These distinctions are not used in our phonemic dictionary, except for one contrastive experiment with soft/hard consonants.

Table 3
Grapheme and phoneme systems.

| System | #Units | Mapping |
|---|---|---|
| FLP-35 graph | 35 | Graphs |
| FLP-32 phone | 32 | Phones |
| FLP-36 phone | 36 | Affricates |
| FLP-38 phone | 38 | Diphthongs, except ou → /ɔʊ/, oi → /ɔi/ |
| FLP-48 phone | 48 | Soft consonants, except soft ch → /x/ |
| VLLP-33 graph | 33 | Graphs, c → /ts/, f → /v/ |
| VLLP-29 graph | 29 | z → /s/, ch → /ɣ/, ę → /ɛ/, į,y → /i:/, ūų → /u:/ |
| VLLP-31 phone | 31 | Phones, f → /v/ |
| VLLP-29 phone | 29 | z → /s/, ch → /ɣ/, ę → /ɛ/, į,y → /i:/, ūų → /u:/ |

In the graphemic lexicon each orthographic character is modeled as a separate grapheme. The rare non Lithuanian characters appearing in the corpus are mapped to Lithuanian ones ($x \rightarrow ks$, $q \rightarrow k$, $w \rightarrow v$).

Linguists do not necessarily agree if affricates and diphthongs should be modeled as a single phoneme or a phoneme sequence. For the 32-phoneme set, affricates are split into a sequence of two phonemes (according to IPA): *c* → */ts/*, *č* → /ʧ/, *dz* → */dz/*, *dž* → */dʒ/*. Soft and hard consonants are represented by the same unit as they can be completely differentiated by their contexts. For the 36-phoneme set, the grapheme-to-phoneme rules represent *c, č dz* and *dž* with single units, which allows the minimal phone duration to be 30 ms instead of 60 ms when the mapping is to a sequence of two phonemes.

To summarize, the FLP-36 and FLP-38 phone sets, respectively, represent affricates and diphthongs as specific units (except for the rare *ou* and *oi*, which were too rare to model and were split into a sequence of vowels). The soft consonants are included in the FLP-48 phone set, with the exception of the rare soft *ch* which is mapped to the hard one.

For the VLLP case, where only 3 h of transcribed speech data are available, several mappings for the poorly modeled rare units were investigated. In all cases the two rarest units (*c* and *f*) are mapped to similar sounding units. Furthermore, in the VLLP-29 graph set and the VLLP-29 phone set, the units *z, ch* and *ę* were mapped because they were rarely observed in the training data. Since the two orthographic pairs *į/y* and *ū/ų* correspond to the same sounds but have different representations due to grammar exceptions, each pair is represented by a single unit.

Speech recognition and keyword search results for five FLP acoustic models, one grapheme-based and four phoneme-based, are shown in Table 4. It can be seen that the explored phoneme mappings have only a slight impact on the number of homophones in the training lexicon. The KWS results were obtained by combining the keyword hits from two systems, one using full-words and the other 7-gram subword character-based units (Hartmann et al., 2014).

Table 5 gives the WER and MTWV results for the VLLP systems. As specified in Section 4, the language models of these VLLP systems are trained on the audio transcripts of the 3 h subset and the distributed Web texts. The acoustic models include semi-supervised training with 77 h of conversation. Both grapheme and phoneme-based models are seen to perform slightly better when the number of units is reduced.

Table 4
WER and MTWV results for the graphemic and phonemic FLP systems listed in Table 3 (top). KWS combines the keyword hits from a full-word system with those of a 7-gram character subword unit system. *Homoph* indicates the number of lexical entries which share a pronunciation.

| System | #Units | Homoph | %WER | MTWV (ALL/INV/OOV) |
|---|---|---|---|---|
| FLP-35 graph | 35 | 522 | 44.6 | 0.579/0.592/0.472 |
| FLP-32 phone | 32 | 719 | 44.7 | 0.576/0.591/0.476 |
| FLP-36 phone | 36 | 718 | 44.6 | 0.580/ 0.593/0.487 |
| FLP-38 phone | 38 | 718 | 44.4 | 0.576/0.591/0.460 |
| FLP-48 phone | 48 | 717 | 44.6 | 0.573/0.587/0.472 |

Table 5
WER and MTWV results for graphemic and phonemic VLLP systems. KWS results from combining the keyword hits from two systems, one full-word system and the other using 7-gram character subword units. *Homoph* indicates the number of lexical entries which share a pronunciation. The VLLP systems used Web texts in the language models and SST for the acoustic models.

| System | #Units | Homoph | %WER | MTWV (ALL/INV/OOV) |
|---|---|---|---|---|
| VLLP-33 graph | 33 | 493 | 52.6 | 0.485/0.496/0.415 |
| VLLP-29 graph | 29 | 1583 | 52.2 | 0.485/0.496/0.417 |
| VLLP-31 phone | 31 | 1336 | 52.3 | 0.491/ 0.504/0.398 |
| VLLP-29 phone | 29 | 2418 | 52.0 | 0.493/0.501/ 0.443 |

It can be observed, that for both the FLP and the VLLP conditions, the different phoneme sets have a limited impact on both speech recognition and keyword search performance.

## 6. Impact of Web data and untranscribed audio

In the above FLP experiments only the manual transcriptions were used for language modeling. To build the VLLP systems, the Web data were also used for training 3-gram language models, and the remaining 77 h of untranscribed data for semi-supervised acoustic model training. These extra resources help to reduce the performance difference between the two conditions. The following experiments aim to assess the impact of the Web data and semi-supervised training for both the FLP and VLLP conditions. These experiments are performed using the 38 phone set for FLP and the 29 phone set for VLLP, as these sets gave the best WER results, and the MTWV results were also best or close to the best.

The STT results assessing the impact of using Web texts for language modeling, and lattice-based semi-supervised acoustic model training (Fraga-Silva et al., 2011) are given in Table 6. Comparing the top rows of each section (FLP 40 h and VLLP 3 h), there is a large difference in the WER obtained by these two systems: with the extra 37 h of audio with transcripts available for the FLP condition, the OOV rate is cut in half and the relative WER reduced by 25% (44.4% vs. 59.3%). As can be expected, the Web texts have a much larger impact for the VLLP condition than for the FLP one: for the VLLP condition the vocabulary is extended from 5.7k to 60k, whereas for the FLP condition the vocabulary size doubles. As a results, there is a much larger reduction in OOVs and WER for the VLLP condition than for FLP. For VLLP the WER is reduced by 6% absolute compared to 2% for FLP. Semi-supervised training does not improve the performance of the FLP system, but does improve the VLLP system performance both with and without Web data.

Table 6
WER results for contrastive training conditions: only manual transcriptions used for LM training, Web texts used, SST used for acoustic modeling. FLP: trn is comprised of 40 h of transcribed speech; VLLP: trn comprised of 3 h of transcribed speech. SST based on 40/77 hours of untranscribed speech for FLP and VLLP, respectively.

| Set | AM | LM | Lexicon | %OOV | %WER |
|---|---|---|---|---|---|
| FLP | trn (40 h) | trn | 30k | 7.6 | 44.4 |
| FLP | trn + SST | trn | 30k | 7.6 | 44.8 |
| FLP | trn | trn + Web | 60k | 5.2 | 42.4 |
| FLP | trn + SST | trn + Web | 60k | 5.2 | 42.4 |
| VLLP | trn (3 h) | trn | 5.7k | 16.7 | 59.3 |
| VLLP | trn + SST | trn | 5.7k | 16.7 | 59.0 |
| VLLP | trn | trn + Web | 60k | 6.0 | 53.3 |
| VLLP | trn + SST | trn + Web | 60k | 6.0 | 52.0 |

## 7. Improving keyword search

Out-of-vocabulary keywords are a challenge for keyword search as they can dramatically affect keyword spotting performance. Various methods have been proposed to address the problem of detecting OOV keywords. One common approach is to convert word lattices to phoneme (or grapheme) lattices and perform phone/grapheme based string search (Siohan and Bacchiani, 2005; Karakos et al., 2014). As proposed in Hartmann et al. (2014), Chaudhari and Picheny (2007), He et al. (2014), lattices can be converted to various-sized subword units. In Chen et al. (2013), the proxy approach is presented, where keyword search allows matches to vocabulary words which are phonetically similar to the specified keyword.

Two approaches to enhance the detection of OOV keywords are employed in this study: using Web texts to increase the lexical coverage and using subword units for keyword search. As the Lithuanian language is highly inflected, it is natural to investigate word morphological decomposition, attempting to automatically derive basic morphological units of the language. Keyword search with morpheme units was reported in Hartmann et al. (2014), Gorin et al. (2015) for such languages as Kazakh, Haitian Creole, Assamese, Bengali, and Zulu. In this work the Morfessor toolkit (Virpioja et al., 2013) is used to extract morphs, applying the non-initial tagging for word-to-morph mapping as described in Gorin et al. (2015). To simplify recombination, all morphs of a word, with the exception of the first unit of each word, are tagged with a special symbol. The 7-gram cross-word character subword units (Hartmann et al., 2014) used to obtain the KWS results presented earlier in Tables 4 and 5 are compared with automatically derived morpheme-based units.

The impact of using Web texts to extend the lexicon and augment the language model on keyword search can be assessed by comparing the upper and lower parts of Tables 7–9. In order to help analyze the impact, performance is reported for different classes of keywords: INV, OOV–INV, OOV–OOV. When words from Web texts are added to the lexicon, OOV words can become INV words (OOV–INV), but some of them will remain OOV (OOV–OOV). The INV keywords are considered with respect to the original lexicon (INV–INV).

Keyword spotting results are given in Table 7 for the FLP condition, and in Tables 8 and 9 for the VLLP condition without and with SST, respectively. For both FLP and VLLP cases, it can be seen that adding Web texts improves the overall MTWV, with the main improvement coming from the better lexical coverage of the augmented LM (row *word*, OOV–INV). However, the remaining OOV keywords (OOV–OOV) are still poorly detected. As can be expected, much larger gains are achieved for the VLLP case than for the FLP due to the dramatic difference in the lexical coverage. As a reminder, Table 6 showed that using the Web texts reduced the OOV rate by 30% for the FLP condition and by 65% for the VLLP one.

Table 7
KWS results in terms of MTWV for FLP systems (no SST). Performance reported using full-words (*word*), character 7-gram subwords (*char-sw*), 7-gram subwords using only OOV hits for combination (*char-sw*$_{oov}$), morpheme subwords (*morph-sw*) and some combinations. Without (upper part) and with (lower part) Web texts.

| Unit | LM | ALL | INV–INV | OOV–INV | OOV–OOV |
|---|---|---|---|---|---|
| Word | trn | 0.544 | 0.596 | – | 0.085 |
| Char-sw | trn | 0.483 | 0.488 | – | 0.445 |
| Morph-sw | trn | 0.509 | 0.521 | – | 0.416 |
| Word+char-sw | trn | 0.576 | 0.591 | – | 0.460 |
| Word+char-sw$_{oov}$ | trn | 0.582 | 0.596 | – | 0.460 |
| Word+morph-sw | trn | 0.571 | 0.590 | – | 0.423 |
| Word+char-sw+morph-sw | trn | 0.573 | 0.585 | – | 0.496 |
| Word+char-sw$_{oov}$+morph-sw | trn | 0.580 | 0.590 | – | 0.496 |
| Word | trn+Web | 0.570 | 0.600 | 0.567 | 0.162 |
| Char-sw | trn+Web | 0.485 | 0.490 | 0.431 | 0.461 |
| Morph-sw | trn+Web | 0.559 | 0.575 | 0.542 | 0.354 |
| Word+char-sw | trn+Web | 0.586 | 0.595 | 0.583 | 0.484 |
| Word+char-sw$_{oov}$ | trn+Web | 0.591 | 0.600 | 0.567 | 0.484 |
| Word+morph-sw | trn+Web | 0.591 | 0.606 | 0.615 | 0.380 |
| Word+char-sw+morph-sw | trn+Web | 0.591 | 0.598 | 0.597 | 0.523 |
| Word+char-sw$_{oov}$+morph-sw | trn+Web | 0.601 | 0.606 | 0.615 | 0.523 |

Table 8
KWS results in terms of MTWV for VLLP systems when only 3 h of transcribed audio are used for training the acoustic models (no SST). Performance measured using full-words (*word*), character 7-gram subwords (*char-sw*), 7-gram subwords using only OOV hits for combination (*char-sw$_{oov}$*), morpheme subwords (*morph-sw*) and some combinations. Without (upper part) and with (lower part) Web texts.

| Unit | LM | ALL | INV–INV | OOV–INV | OOV–OOV |
|---|---|---|---|---|---|
| Word | trn | 0.271 | 0.445 | – | 0.034 |
| Char-sw | trn | 0.353 | 0.361 | – | 0.344 |
| Morph-sw | trn | 0.347 | 0.403 | – | 0.281 |
| Word+char-sw | trn | 0.401 | 0.450 | – | 0.342 |
| Word+char-sw$_{oov}$ | trn | 0.401 | 0.445 | – | 0.342 |
| Word+morph-sw | trn | 0.373 | 0.454 | – | 0.279 |
| Word+char-sw+morph-sw | trn | 0.412 | 0.454 | – | 0.372 |
| Word+char-sw$_{oov}$+morph-sw | trn | 0.416 | 0.454 | – | 0.372 |
| Word | trn+Web | 0.444 | 0.485 | 0.491 | 0.152 |
| Char-sw | trn+Web | 0.375 | 0.371 | 0.373 | 0.405 |
| Morph-sw | trn+Web | 0.436 | 0.458 | 0.456 | 0.314 |
| Word+char-sw | trn+Web | 0.477 | 0.486 | 0.497 | 0.417 |
| Word+char-sw$_{oov}$ | trn+Web | 0.474 | 0.485 | 0.491 | 0.417 |
| Word+morph-sw | trn+Web | 0.476 | 0.499 | 0.511 | 0.318 |
| Word+char-sw+morph-sw | trn+Web | 0.483 | 0.493 | 0.508 | 0.443 |
| Word+char-sw$_{oov}$+morph-sw | trn+Web | 0.492 | 0.499 | 0.510 | 0.443 |

In all conditions, subword units are seen to detect a significant portion of the OOV keywords. Comparing 7-gram character subwords (row *char-sw*) with the morpheme subwords (column *morph-sw*), it can be seen that the performance with the former is better for OOV–OOV, and less good for INV–INV and OOV–INV. This difference in performance is larger when the Web texts are used.

Various combinations of full-word and subword based systems were evaluated in an attempt to reap the benefits of both approaches. It can be seen that in some cases combining keyword hits obtained with word and 7-gram subword units (entries word vs. word+char-sw) results in a small degradation of performance on the INV keywords. This loss is eliminated by combining the word hits with hits produced by the 7-gram subword system only for the OOV keywords (row *char-sw$_{oov}$*). Combining the hits produced with the full-word systems with those produced with the morph-based subword units always improves performance over either approach alone.

Comparing the *word* entries in the upper and lower parts of Table 7 it can be seen that the Web texts improve the MTWV from 0.544 to 0.570 for the FLP system. This gain is largely attributed to the detection of OOV keywords

Table 9
KWS results in terms of MTWV for VLLP systems using SST for acoustic modeling. Performance measured using full-words (*word*), character 7-gram subwords (*char-sw*), 7-gram subwords using only OOV hits for combination (*char-sw$_{oov}$*), morpheme subwords (*morph-sw*) and some combinations. Without (upper part) and with (lower part) Web texts.

| Unit | LM | ALL | INV–INV | OOV–INV | OOV–OOV |
|---|---|---|---|---|---|
| Word | trn | 0.288 | 0.466 | – | 0.046 |
| Char-sw | trn | 0.377 | 0.384 | – | 0.370 |
| Morph-sw | trn | 0.371 | 0.414 | – | 0.314 |
| Word+char-sw | trn | 0.420 | 0.466 | – | 0.371 |
| Word+char-sw$_{oov}$ | trn | 0.425 | 0.466 | – | 0.371 |
| Word+morph-sw | trn | 0.400 | 0.468 | – | 0.315 |
| Word+char-sw+morph-sw | trn | 0.431 | 0.465 | – | 0.403 |
| Word+char-sw$_{oov}$+morph-sw | trn | 0.436 | 0.468 | – | 0.403 |
| Word | trn+Web | 0.466 | 0.501 | 0.512 | 0.201 |
| Char-sw | trn+Web | 0.399 | 0.394 | 0.404 | 0.417 |
| Morph-sw | trn+Web | 0.453 | 0.469 | 0.483 | 0.336 |
| Word+char-sw | trn+Web | 0.493 | 0.498 | 0.512 | 0.443 |
| Word+char-sw$_{oov}$ | trn+Web | 0.495 | 0.501 | 0.512 | 0.443 |
| Word+morph-sw | trn+Web | 0.488 | 0.506 | 0.532 | 0.343 |
| Word+char-sw+morph-sw | trn+Web | 0.497 | 0.497 | 0.523 | 0.452 |
| Word+char-sw$_{oov}$+morph-sw | trn+Web | 0.504 | 0.506 | 0.532 | 0.452 |

that become INV. Their detection is seen to be close to that of the original INV keywords, for which acoustic examples were observed in the training corpus. Combining the keyword hits produced by full-word search and the two subword systems results in an overall MTWV of 0.601 and improves the detection of OOV–OOV keywords by 36.1 points (0.162 vs. 0.523) over the full-word system.

As previously noted, using the Web texts has a larger impact on the VLLP results (cf. Tables 8 and 9) than the FLP ones. With word-based keyword search the MTWV is increased by over 17 points (from 0.271 to 0.444 without SST, and from 0.288 to 0.466 with SST).

As observed for the FLP condition, for VLLP combining of keyword hits improves the OOV–OOV keyword detection compared to full-word search alone. Using the Web texts, the OOV–OOV detection is improved by 29.1 points absolute (0.152 vs. 0.443) without SST, and 25.1 points (0.201 vs. 0.452) with SST.

Comparing the detection results with full-word search to the two subword-based approaches, the 7-gram character subword search obtains the best performance on out-of-vocabulary keywords, whereas the full-word search consistently obtains the best results on in-vocabulary keywords. The morpheme-based subword units produce better results on in-vocabulary keywords and less good results on out-of-vocabulary keywords than the 7-gram character-based units.

## 8. Summary

This paper has reported on research carried out to develop systems for transcription and keyword search in conversational telephone speech for the low-resourced Lithuanian language. According to the linguistic literature, the phonemic inventory for Lithuanian is generally described with 56 phonemes. However, when resources are limited, some of the phonemes may not be sufficiently (or at all) observed. Experiments were carried out with different phoneme inventories to determine the best set of units to model, more specifically to assess the impact of explicitly modeling diphthongs, affricates and soft consonants. The different phoneme inventories explored had a limited impact on both speech recognition and keyword search performance. Since developing a pronunciation dictionary requires human expertise, phoneme- and grapheme-based acoustic models were compared. As has been reported using grapheme-based acoustic units for other languages, the phoneme models gave only a slight improvement for the two training conditions (3 or 40 h of transcribed audio data). We attribute this small difference to the strong dependency between the orthographic and phonemic forms in Lithuanian.

The impact of using Web texts for training language models, and untranscribed data for semi-supervised training of the acoustic models was also assessed. Adding Web texts in the FLP language model and expanding the vocabulary to 60k words reduced the WER by almost 2% absolute, but SST was not helpful. For the VLLP condition, the WER was reduced by more than 7% absolute using both the Web data (with a 60k word vocabulary) and semi-supervised training.

Keyword search was carried out using words and two types of automatically derived subword units. Word-based search obtains the best results for in-vocabulary keywords, but has very poor performance for out-of-vocabulary keywords. Comparing the two types of subwords, the 7-gram character subwords led to better performance on out-of-vocabulary keywords and worse on in-vocabulary ones than the morpheme-based subwords. The best keyword spotting results are achieved using the Web data to augment the language models and combining hits produced using full-word, character-subword, and morpheme-subword based search.

## 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.

## Acknowledgments

We would like to thank our IARPA Babel partners for sharing resources (BUT for the bottleneck features and BBN for the Web data), and Grégory Gelly for providing the voice activity detector.

This research was in part supported by the French National Agency for Research as part of the SALSA (Speech And Language technologies for Security Applications) project under grant ANR-14-CE28-0021, and 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.

## References

Adda-Decker, M., Lamel, L., 2000. The use of Lexica in automatic speech recognition. In: Proceedings of the 2009 Lexicon Development for Speech and Language Processing. Springer, pp. 235–266.

Chaudhari, U.V., Picheny, M., 2007. Improvements in phone based audio search via constrained match with high order confusion estimates. In: Proceedings of the 2007 Automatic Speech Recognition & Understanding (ASRU), pp. 665–670.

Chen, G., Yilmaz, O., Trmal, J., Povey, D., Khudanpur, S., 2013. Using proxies for OOV keywords in the keyword search task. In: Proceedings of the 2013 Automatic Speech Recognition & Understanding (ASRU), pp. 416–421.

Cui, J., Mamou, J., Kingsbury, B., Ramabhadran, B., 2014. Automatic keyword selection for keyword search development and tuning. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7839–7843.

Filipovič, M., Lipeika, A., 2004. Development of HMM/neural network-based medium-vocabulary isolated-word Lithuanian speech recognition system. Informatica 15 (4), 465–474.

Fiscus, J.G., Ajot, J., Garofolo, J.S., Doddington, G., 2007. Results of the 2006 spoken term detection evaluation. In: Proceedings of the 2007 International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Vol. 7, pp. 51–57.

Fraga-Silva, T., Gauvain, J.L., Lamel, L., 2011. Lattice-based unsupervised acoustic model training. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4656–4659.

Gales, M., Knill, K., Ragni, A., 2015. Unicode-based graphemic systems for limited resource languages. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5186–5190.

Gauvain, J.L., Lamel, L., Adda, G., 2002. The LIMSI broadcast news transcription system. Speech Communication 37 (1), 89–108.

Gelly, G., Gauvain, J.L., 2015. Minimum word error training of RNN-based voice activity detection. In: Proceedings of the 2015 Conference of the International Speech Communication Association (INTERSPEECH), pp. 2650–2654.

Girdenis, A., 2003. Teoriniai Lietuvių Fonologijos Pagrindai. Mokslo ir enciklopedijų leidybos institutas.

Gorin, A., Lamel, L., Gauvain, J.L., Fraga-Silva, T., 2015. On improving speech recognition and keyword spotting with automatically generated morphological units. In: Proceedings of the 2015 Conference on Language and Technology Conference (LTC).

Grézl, F., Karafiát, M., 2014. Combination of multilingual and semi-supervised training for under-resourced languages. In: Proceedings of the 2014 Conference of the International Speech Communication Association (INTERSPEECH), pp. 820–824.

Harper, M., 2013. The BABEL program and low resource speech technology. In Automatic Speech Recognition & Understanding Workshop (ASRU) Invited talk. https://www.iarpa.gov/index.php/research-programs/babel/baa, (Accessed in 2017).

Hartmann, W., Le, V.B., Messaoudi, A., Lamel, L., Gauvain, J.L., 2014. Comparing decoding strategies for subword-based keyword spotting in low-resourced languages. In: Proceedings of the 2014 Conference of the International Speech Communication Association (INTERSPEECH), pp. 2764–2768.

He, Y., Hutchinson, B., Baumann, P., Ostendorf, M., Fosler-Lussier, E., Pierrehumbert, J., 2014. Subword-based modeling for handling OOV words in keyword spotting. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7864–7868.

Kanthak, S., Ney, H., 2002. Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2, pp. 845–848.

Karakos, D., Bulyko, I., Schwartz, R., Tsakalidis, S., Nguyen, L., Makhoul, J., 2014. Normalization of phonetic keyword search scores. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7834–7838.

Karakos, D., Schwartz, R., Tsakalidis, S., Zhang, L., Ranjan, S., Tim Ng, T., Hsiao, R.C., Saikumar, G., Bulyko, I., Nguyen, L., et al., 2013. Score normalization and system combination for improved keyword spotting. In: Proceedings of the 2013 Automatic Speech Recognition & Understanding (ASRU), pp. 210–215.

Kazlauskienė, A., Raškinis, G., 2009. Bendrinės Lietuvių kalbos garsų dažnumas. Respect. Philol. 16 (21), 169–182.

Kemp, T., Waibel, A., 1999. Unsupervised training of a speech recognizer: recent experiments. In: Proceedings of the 1999 European Conference on Speech Communication and Technology (Eurospeech), pp. 2725–2728.

Killer, M., Stüker, S., Schultz, T., 2003. Grapheme based speech recognition. In: Proceedings of the 2003 Conference of the International Speech Communication Association (INTERSPEECH).

Lamel, L., 2013. Unsupervised acoustic model training with limited linguistic resources. In: Proceedings of the 2013 Automatic Speech Recognition & Understanding (ASRU).

Lamel, L., Gauvain, J.L., Adda, G., 2002. Lightly supervised and unsupervised acoustic model training. Comput. Speech Lang. 16 (1), 115–129.

Laurinčiukaitė, S., Lipeika, A., 2015. Syllable-phoneme based continuous speech recognition. Elektron. Elektrotech. 70 (6), 91–94.

Lileikyte, R., Lamel, L., Gauvain, J.L., 2015. Conversational telephone speech recognition for Lithuanian. In: Proceedings of the 2015 International Conference on Statistical Language and Speech Processing (SLSP), pp. 164–172.

Lileikytė, R., Lamel, L., Gauvain, J.L., Gorin, A., 2016. Lithuanian broadcast speech transcription using semi-supervised acoustic model training. In: Proceedings of the 2016 Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU).

Lipeika, A., Lipeikienė, J., Telksnys, L., 2002. Development of isolated word speech recognition system. Informatica 13 (1), 37–46.
Mangu, L., Brill, E., Stolcke, A., 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Comput. Speech Lang. 14 (4), 373–400.
Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V., 2015. Investigation of foreign languages models for Lithuanian speech recognition. Elektron. Elektrotech. 91 (3), 15–20.
Pakerys, A., 2003. Lietuvių Bendrinės Kalbos Fonetika. Enciklopedija.
Prasad, R., Matsoukas, S., Kao, C.L., Ma, J.Z., Xu, D., Colthurst, T., Kimball, O., Schwartz, R.M., Gauvain, J.L., Lamel, L., et al., 2005. The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system. In: Proceedings of the 2005 Conference of the International Speech Communication Association (INTERSPEECH), pp. 1645–1648.
Raškinis, G., Raškinienė, D., 2003. Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. Informatica 14 (1), 75–84.
Šilingas, D., Laurinčiukaite, S., Telksnys, L., 2004. Towards acoustic modeling of Lithuanian speech. In: Proceedings of the 2004 International Conference on Speech and Computer (SPECOM), pp. 326–333.
Šilingas, D., 2005. Akustinių Lietuvių Šnekos Atpažinimo Modelių Parinkimas, Naudojant Paslėptus Markovo Modelius, Ph.D. thesis, Vytautas Magnus University.
Siohan, O., Bacchiani, M., 2005. Fast vocabulary independent audio search using path-based graph indexing. In: Proceedings of the 2005 Conference of the International Speech Communication Association (INTERSPEECH), pp. 53–56.
Vaičiūnas, A., Raškinis, G., 2006. Cache-based statistical language models of English and highly inflected Lithuanian. Informatica 17 (1), 111–124.
Vaišnienė, D., Zabarskaitė, J., Rehm, G., Uszkoreit, H., 2012. The Lithuanian Language in the Digital Age. Springer.
Virpioja, S., Smit, P., Grönroos, S.A., Kurimo, M., et al., 2013. Morfessor 2.0: Python 2.0: implementation and extensions for Morfessor baseline. Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Aalto University. Finland.
Zavaliagkos, G., Colthurst, T., 1998. Utilizing untranscribed training data to improve performance. In: Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop, pp. 301–305.
Zhang, L., Karakos, D., Hartmann, W., Hsiao, R., Schwartz, R., Tsakalidis, S., 2015. Enhancing low resource keyword spotting with automatically retrieved web documents. In: Proceedings of the 2015 Conference of the International Speech Communication Association (INTERSPEECH).