Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

© 2014 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-442-8-75

Baseline for Keyword Spotting in Latvian Broadcast Speech

Roberts DARĢIS[a,1] and Artūrs ZNOTIŅŠ[a]
[a]*Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd., Riga, Latvia*

Abstract. This paper describes the first efforts of keyword identification in unconstrained Latvian broadcast speech. During this research a large vocabulary continuous speech recognition (LVCSR), spotting in LVCSR lattices and acoustic keyword spotting had been compared. Open source tools and recently created 100 hours of Latvian Speech Recognition Corpus have been used.

Keywords. Acoustic keyword spotting, LVCSR, Lattices

Introduction

Keyword spotting (KWS) is the task of detecting keywords of interest within continuous speech. It is used in a variety of applications including data mining and audio document indexing. There are multiple methods used for keyword spotting.

In this paper, we present a baseline for KWS in Latvian using available open source tools and 100 hours of audio corpus. For experiments, we used *CMU Sphinx*[2] that is an open source toolkit for speech recognition.

For our purposes the main requirements of KWS system were as follows:

- Primary language is Latvian, though foreign words (mainly organization names) can occur.
- System should support thousands of searched keywords since Latvian is an inflectional language.
- Real time factor should not be more than 1.0 on a single core CPU.

1. Evaluation

In the definition of keyword set, we have selected 244 keywords (1,520 inflected forms) of interest that includes location, product and organization names that occur in evaluation data. Searched keywords also include multiword expressions (e.g., *Latvian National Opera*, *Ministry of Foreign Affairs*). Statistics about chosen keyword set are shown in Figure 1.

[1] Corresponding Author: Roberts Dargis, Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd., Riga, Latvia; E-mail: roberts.dargis@lumii.lv

[2] Available: http://cmusphinx.sourceforge.net

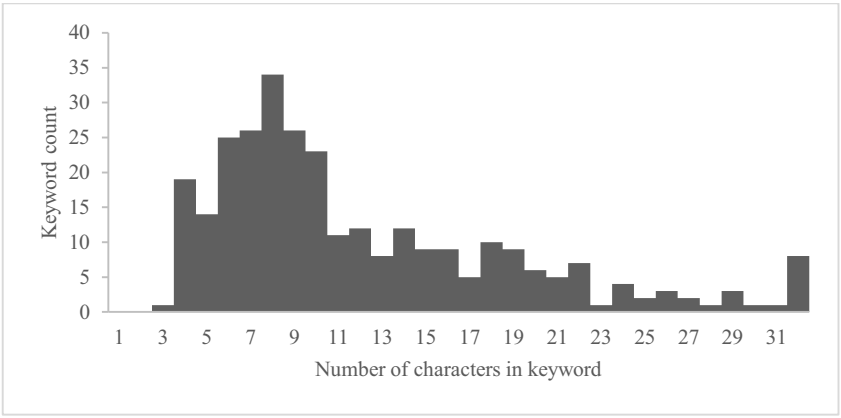

Figure 1. Distribution of keyword lengths in the test set.

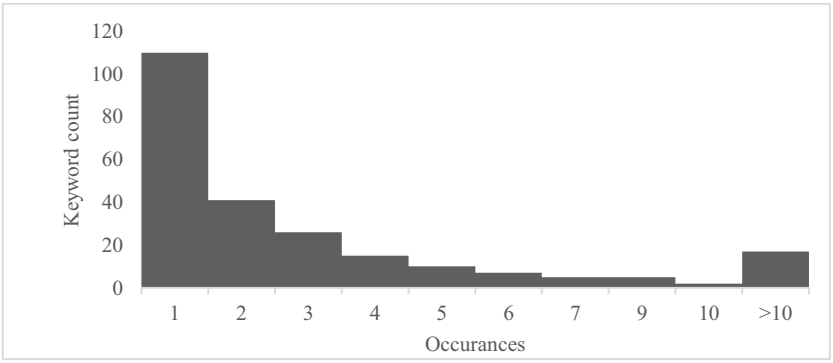

Figure 2. Count of keywords by occurrences in the test set

For evaluation purposes, 2.5 hours audio data set collected from various Latvian broadcasts was used, as well as speaker-independent speech recognition with audio data from 40 different speakers. There are approximately 1,100 mentions of keywords, roughly one mention in every 8 seconds. The average length of keyword in test data set is 9.34 characters and 1.22 words. All audio files were split in the first silent after at least 15 seconds. Keyword is assumed to be recognized correctly if it is present in gold transcription of an audio file fragment. Two evaluation settings were used: precise (keyword form must be recognized precisely) and relaxed (mismatch of inflected forms is allowed). We also evaluated nested keywords (smaller keyword is a part of a larger keyword), but these results do not differ significantly from taking into account only the longest keyword.

For our experiments, we evaluated precision, recall and overall accuracy using F1-score.

2. LVCSR KWS

By using large vocabulary continuous speech recognition, keyword spotting was implemented. It consisted of two stages: full text recognition[3] and text-based search to locate the keywords.

In the first stage *PocketSphinx* engine (*Sphinx 4* module gives similar results) finds the most probable sequence of words based on the Viterbi search algorithm, using an acoustic model, phonetic dictionary and language model. In the second stage, KWS uses LVCSR output using text-based search to locate the keywords.

It is limited to tune balance between recall and precision by using only the most probable output of LVCSR. For this reason, experiments searching for keywords in LVCSR lattices that are useful for analyzing alternative hypothesis were carried out. Lattice is a directed graph that contains nodes representing words spoken over a particular period of time and edges that correspond to the score of one word following another. Although we discovered that branching factor of lattices is rather hard to tune and it highly increases processing time, this method does improve results. Searching for keywords in lattice graph significantly increases recall but also decreases precision. LVCSR recognizer assigns each word acoustic and language model likelihoods. Using forward-backward inference algorithm we computed forward, backward and posterior scores. Using the same confidence scoring technique as [1] we assigned each word with a confidence score (1) that is computed by the forward likelihood $L_\alpha(N)$ of the best path through lattice from the beginning of lattice to the keyword, backward likelihood $L_\beta(N)$ from the end of lattice, word posterior L(N) and best path through lattice L_{best}:

$$C(N) = L_\alpha(N) + L_\beta(N) + L(N) - L_{best} \tag{1}$$

$$L_\alpha(N) = L_{acoustic}(N) + L_{language}(N) + \min_{N_P} L_\alpha(N_P) \tag{2}$$

$$L_\beta(N) = L_{acoustic}(N) + L_{language}(N) + \min_{N_F} L_\beta(N_F), \tag{3}$$

where $L_{acoustic}$ is acoustic model score, $L_{language}$ is language model score.

As keyword may contain more than one word, the largest confidence score for word within the keyword was chosen as a whole keyword confidence score. The performance of the final system was tuned by computed confidence score. Real time factor was about 0.5.

3. Acoustic KWS

Acoustic KWS uses an acoustic model and a phonetic dictionary that contain pronunciations of searched keywords. Simple background and filler models are used to model non-keyword speech.

[3] Demo for LVCSR available http://85.254.250.60/speech_recognition/

For acoustic KWS experiment *Sphinx4* long audio aligner implementation was used. It contains simplistic keyword grammar that consists of all the keywords (in parallel) that are to be spotted (and only those). Out of grammar words were modeled with phone loops right at the beginning of words in the grammar. Multiword keywords were simply concatenated. Performance of the system was tuned by 3 parameters: phoneme insertion, word insertion and out of grammar probability, (see Figure 3). Configurations with relatively high probability of branching out to phoneme loops give the best results.

Although this kind of method can give good and fast results [2], nowadays it is not used so often. Processing time increased linearly by adding new keywords to the used *Sphinx4* model dictionary (2,000 keyword entries increased real time factor to 1.0). As the initial results were not so promising we concentrated more on LVCSR based KWS.

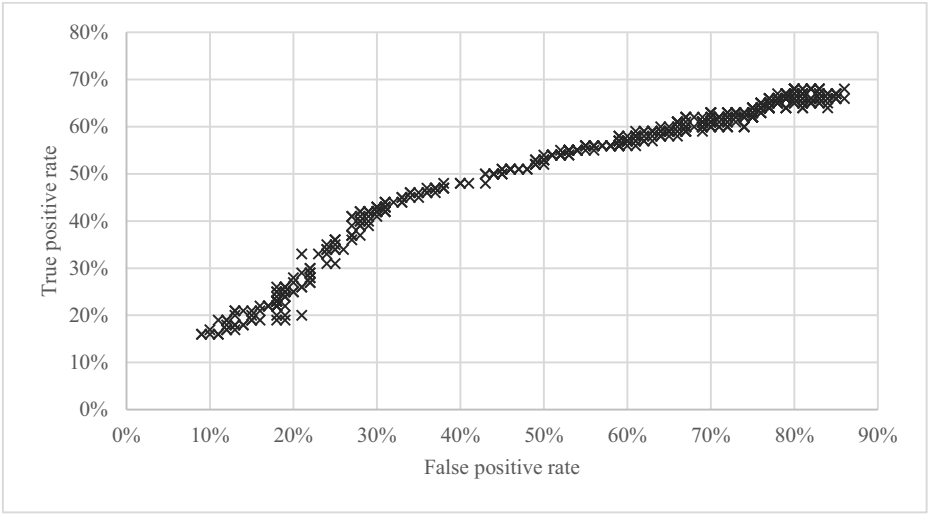

Figure 3. Receiver operating characteristic (ROC) curve for acoustic KWS.

4. Resources

4.1. Dictionary

Latvian is highly inflective language. Numerous lemmas have more than six different corresponding word-forms. Some of the word-forms might never be used in any reasonable sentence construction. To avoid including these forms in our dictionary, instead of inflecting all words from Latvian dictionary, we obtained all unique text tokens from a large set of Latvian news articles, containing more than 175 million tokens. At the end, dictionary consisted of almost 600 thousand word-forms.

Phonetic transcription was generated by a rule based system that uses approximately 250 expert defined rules and 1,300 exceptions. This system was initially built for speech synthesis thus phoneme set was more detailed than necessary. We managed to reduce phoneme set from 68 to 57 phonemes. Suitability of phonetic transcription for informal speech had been evaluated [3].

4.2. Acoustic Model

The acoustic model is trained on 100 hours audio data from the Latvian Speech Recognition Corpus. It has been designed to represent the major speech characteristics of Latvian population [4]. The main source of corpus audio data is the same as our target audio data: Latvian TV and radio broadcasts.

The CMU Sphinx toolkit was used to develop an acoustic model. The AM is context-depended continuous triphone HMM with 4,000 tied states.

The AM contains 57 phoneme models, a silence model and 6 different noise models. Noise models include hesitation, loud inhalation or exhalation and physiological noises (cough, laughter).

For feature extraction from 16bit 44,100 kHz audio, we used 13 dimensional feature vectors.

4.3. Language Model

Text source is the same as training text data from the acoustic model. Text contains 1M tokens with 70k unique word-forms. Language model vocabulary is limited to word-forms that occurred more than once, manually adding inflected keywords. We used open vocabulary model with OOV factor of 0.3 and the CMU-Cambridge Statistical Language Modeling Toolkit v2[4] for language modeling.

5. Results

5.1. Overall Results

Initial results revealed that used acoustic KWS produces higher false alarm ratio and it is hard to tune it to reach a better results. First, because it does not use language model information and second, because different keywords are similar to other words or parts of words in varying degree. Keyword length is also important (longer words are recognized more precisely).

Processing time used for acoustic KWS increases linearly by adding new keywords. Real time results are possible only if keyword vocabulary contains less than 1,000 keyword forms (this introduces the problem of optimizing vocabulary for larger keyword set). This means that we should optimize vocabulary by using stems instead of all keyword forms.

Although lattice searching and scoring does not lead to a major increase in performance of the used toolkit, it does help to fine tune and find balance between precision and recall.

Initial results (Table 1) revealed that LVCSR system leads to much better results. Related work [1] has shown that LVCSR and acoustic KWS should give similar results. This means that used Sphinx-4 acoustic module is not suitable for rather large keyword vocabulary KWS.

[4] Available: http://www.speech.cs.cmu.edu/SLM/toolkit.html

Table 1. Results.

	F1	Precision	Recall	True positive	False positive	False negative
Precise						
LVCSR	0.67	0.76	0.59	743	232	516
Lattice searching	0.64	0.66	0.63	787	397	472
Lattice scoring	0.67	0.74	0.61	770	265	489
Acoustic	0.37	0.45	0.32	401	494	858
Relaxed						
LVCSR	0.78	0.88	0.69	870	115	389
Lattice searching	0.77	0.80	0.75	939	237	320
Lattice scoring	0.81	0.90	0.73	923	108	336
Acoustic	0.47	0.56	0.40	503	401	756

The recognition results shows:

- the number of correctly recognized keywords (true positive);
- the number of falsely recognized keywords (false positive);
- the number of missed keywords (false negative).

We used precision (4) to measure how many of recognized keywords were identified correctly and we used recall (5) to measure how many of keywords in test data were not missed in recognition. To compare the overall accuracy of different systems, we used F1 score (6) (also called harmonic mean).

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{4}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{5}$$

$$F1 = \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

5.2. WER Breakdown by Quality of Recordings

One of the major factors that affect quality of KWS in all our explored methods is the quality of the audio recordings. To further assess the impact, we used LVCSR and measured full text recognition word error rate (WER) (7) and accuracy (8).

$$WER = \frac{substitutions + insertions + deletions}{reference\ length} \tag{7}$$

$$accuracy = \frac{correct\ words}{reference\ length} \tag{8}$$

WER, accuracy and percentage of data are shown in Table 2. The best accuracy is achieved with clean studio recordings. Recordings with loud background noise (for

example, street noise or music) or with poor recording quality (for example, telephone conversation) rise significantly lower accuracy. Noise filtering before KWS might increase accuracy in noisy recordings (12.24% of data).

Table 2. Full text recognition quality and percentage of data by type of recordings.

Type of recording	WER	Accuracy	Percentage of data
Outside studio with background noise	46.55%	58.14%	15.68%
Studio with background noise	47.07%	57.76%	43.46%
Studio	47.56%	60.58%	27.28%
Outside studio without background noise	48.44%	57.03%	1.34%
On street	52.73%	52.73%	1.17%
Studio with background music	56.13%	48.31%	6.98%
Telephone	87.47%	24.30%	4.09%

5.3. LVCSR Based KWS Error Analysis

21.9% of errors are related with incorrectly recognized keyword inflection. For multiword keywords, these errors are mainly caused by incorrectly recognized case of the last word. In general, first words of multiword expressions are used in the genitive case that can be easily identified by a language model.

False negatives (51.1%) are caused by incorrectly recognized words (50%), words or parts of words similar in the reference text (29%), some of the words recognized incorrectly (21%). 16% of these errors are located in the beginning of audio files. False positives (27.0%) give similar error group distribution to false negatives (46%, 30% and 24% respectively).

In conclusion, main cause of the errors is caused by a rather large word error rate and incorrectly recognized words or similar words.

6. Conclusion

In this paper, we have compared several methods for keyword spotting. Baseline results have shown that the best approach for large vocabulary keyword spotting is continuous speech recognition with the possibility to tune between precision and recall using search in lattices. This will be our main approach in our further work with the following aims:

- add noise filtering;
- implement factored language model for more precise word form recognition;
- configure separate threshold values for keywords.

7. Acknowledgements

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Center"[5] of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 2.10 "Exploring the potential of automated speech recognition for Latvian media monitoring".

References

[1] I. Szöke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapso, J. Cernocký, Comparison of keyword spotting approaches for informal continuous speech. In *Interspeech*, 2005, 633–636.

[2] J. Nouza, J. Silovský, Fast Keyword Spotting in Telephone Speech. In *Radioengineering*, 2009, 665–670.

[3] Auziņa, I., Pinnis, M., & Darģis, R. (2014). Comparison of Rule-based and Statistical Methods for Grapheme to Phoneme Modelling. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania: IOS Press.

[4] Pinnis, M., Auziņa, I., & Goba, K. (2014). Designing the Latvian Speech Recognition Corpus. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

[5] Information and Communication Technology Competence Center: http://www.itkc.lv/