Keyword spotting for audiovisual archival search in Uralic languages

Nils Hjortnæs
Indiana University
nhjortn@iu.edu

Niko Partanen
University of Helsinki
Helsinki, Finland
niko.partanen@helsinki.fi

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Abstract

In this study we investigate the potential of using Automatic Speech Recognition (ASR) for keyword spotting for four Uralic languages: Finnish, Hungarian, Estonian and Komi. These languages also represent different levels on the high and low resource continuum. Although the accuracy of the ASR systems show there is a long way to go, we show that they still have potential to be useful for downstream tasks such as keyword spotting. By using a simple text search after running ASR, we are already able to achieve an F_1 score of between 0.15 and 0.33, a precision of nearly 0.90 for Estonian and Hungarian, and a precision of 0.76 for Komi.

Tiivistelmä

Tutkimus käsittelee puheentunnistuksen käyttöä avainsanojen tunnistamisessa neljällä uralilaisella kielellä, joita ovat suomi, unkari, viro ja komisyrjääni. Nämä kielet ovat myös eri tasoilla saatavilla olevien resurssien määrän suhteen. Vaikka varsinaiset puheentunnistusjärjestelmät eivät välttämättä vielä toimi toivotulla tavalla, osoitamme, että näitä teknologioita voi jo hyödyntää eri tehtävissä, joista yksi on avainsanojen tunnistus. Kokeissamme avainsanat tunnistetaan suoraan puheentunnistuksen tuottamasta tekstistä. Näin saavutettu tarkkuus on verrattain korkea, mutta herkkyys yhä melko matala.

1 Introduction

Very large quantities of audio recordings exist for Uralic languages, as there is a long history of primary data collection. It is another question how large a portion of these materials are adequately archived, and if they are, whether they are findable and accessible. The situation is continuously improving, and as different archives digitize their collections, the material that can be used relatively easily will keep increasing in size. At the same time materials that are not transcribed, translated or annotated can be very challenging to work with. This problem is not unique to the Uralic language materials, nor linguistic materials in general, but touches archived data very widely.

Computational methods have been recognized as one approach to this issue, and several of the related technologies already give very good results (Blokland et al., 2019). When it comes to speech data, it still remains a challenge to develop high performance speech recognition for endangered or low-resource languages (Xu et al., 2020; Stoian et al., 2020). There has, however, been continuous progress in this field to build tools and methods that would allow integration of speech recognition technology into language documentation workflows (see i.e. Adams et al., 2020).

In this study we investigate the usability of using Automatic Speech Recognition (ASR) for keyword spotting for four Uralic languages: Finnish, Hungarian, Estonian and Komi. This way even ASR models that currently have a lower accuracy could be used effectively in some downstream tasks, of which keyword spotting is an important one. For example, there are often recordings that have accompanying notes or metadata, from which potential keywords can be extracted. In long recordings, locating these sections is, however, very tedious and slow to conduct manually. Keyword spotting would allow easier navigation and verification work with unannotated recordings.

2 Wider context of archived multimedia

To contextualize even partly how large the scale of unannotated but existing multimedia is, we use Komi as the example in this section. Focus to Komi in this section is also motivated by the fact that the venue where our study is published is at Syktyvkar, Komi Republic, and Komi is the only endangered language which we address, and thereby the need to accurately locate Komi materials also more urgent. We are most familiar with the European archives, and focus to those, although most substantial Komi collections certainly are stored in Syktyvkar. The first audio recordings of Permian Komi and Udmurt were most likely done in 1911 (Денисов, 2014, 34), which is now 110 years ago. This tells that the materials have accumulated for a long period already.

The Archive of Estonian Dialects and Finno-Ugric Languages at the Institute of the Estonian Language (Ermus et al., 2019) contains a large number of recordings in various Uralic languages, and their online catalogue lists 212 Komi recordings that total in 19 hours. Most of their Komi materials have been collected by Anu-Reet Hausenberg and Adolf Turkin.

Similarly, the Institute for the Languages of Finland contains large Komi collections. These start with the work of Erkki Itkonen, who did a fieldwork trip to Syktyvkar in 1958 an (Itkonen, 1958, 70). Very soon after this Günter Johannes Stipa conducted similar trip (Stipa, 1962, 65–66). We also have to highlight the collections Muusa Vahros-Pertamo did in 1962 both with Zyrian and Permian Komi dialects (Vahros-Pertamo, 1963). These materials have not been published. In 1950s and 1960s Erik Vászolyi conducted similar work, and his recordings were later published (Vászolyi-Vasse, 1999), but also copied by Pertti Virtaranta to Helsinki. Also the recordings of Vászolyi do contain several hours of unpublished materials, primarily conversations. The case of Vászolyi is particularly interesting, as the same recordings must be currently copied in several locations: Helsinki, Syktyvkar, Budapest and Perth, Australia, where he was last located before his death. These recordings are approximately 20 hours.

3 Related work

Speech recognition has been previously studied on all of these languages, and some earlier work on keyword spotting also exists. For Finnish and Estonian ASR technologies have already been developed for a long period of time. Among the most recent studies in Finnish ASR is Jain et al. (2020), and for Estonian Alumäe et al. (2019). Enarvi et al. (2017) addressed both of these languages at the same time. A common point of research has been the need to address sub-word segmentation in various ways, as the agglutinative structure of these languages makes the number of unseen word forms potentially very high. At the same time, when the models have been trained with data from media broadcasts and parliamentary proceedings, the recognition of various conversational genres remains a challenge. Work on keyword spotting, or document retrieval in general, has been more scarce, but (Turunen and Kurimo, 2008) have studied the detection of morphemes from unsegmented Finnish audio recordings.

Several experiments for Komi ASR have been conducted, but the quality has not yet reached levels where the models are particularly useful. The steady progress the work has yielded, however, warrants optimism. In the first reported experiment the results were extremely bad, but demonstrated that in principle these systems can be trained with the currently available data, and some insight was shown to the roles the language models and transfer learning may have in the training process (Hjortnaes et al., 2020). A later study refined the language model with online materials, which improved the result considerably (Hjortnaes et al., 2020). All these models used English as the source language in transfer learning. Most recently an investigation was done about the possible use of other languages, and the transfer learning with Russian Common Voice data was tested (Hjortnæs et al., 2021). The results improved due to changes in the DeepSpeech architecture between different versions, but the English transfer learning still gave better results due to the quantity of data available. Further testing of these models by the authors has shown that producing an accurate transcript from a very clearly pronounced Komi speech can work relatively well. In real spontaneous speech the results are extremely sporadic. However, since there is also a clear ratio of correctly recognized words, or their parts, we believe testing the model in real world scenarios for other down stream tasks such as keyword spotting could be very beneficial. When we search for words we expect to occur in the text, we ignore the impact of entirely incorrectly recognized words, and by boosting the individual keywords we improve the possibility of recognizing the words we want to find

even further. Unfortunately this scenario is not entirely realistic, as in many instances we cannot know what themes and words are present. However, there are also many instances where metadata containing keyword and topic information exists, and the researchers who have done the recordings often have acute information about the topics covered, which they may want to locate in the recordings more automatically.

Within the research of ASR at Uralic languages we can also mention the study on Samoyedic languages by Partanen et al. (2020), where relatively good accuracies were reported for single speaker scenarios. In the context of minority languages spoken in Russia, Wisniewski et al. (2020) also reported recently on their experiment with Bashkir. There have also been approaches to create keyword spotting without an ASR system at the background (van der Westhuizen et al., 2021).

4 Test data

In the test data we look at two compendia. The first is the Common Voice (Ardila et al., 2020) collection of the data for Hungarian and Estonian, and the second is the collection of available data for Finnish and Komi. The datasets are described below, with the first selection representing more artificial read literary language sentences, and the second containing spontaneous spoken language.

4.1 Common Voice

Common Voice (Ardila et al., 2020) is a project aimed at collecting speech data for all of the world's languages. One of the advantages of Common Voice is that, for the languages supported, it provides a very convenient way to contribute and distribute voice recordings. The data consists of short sentences, typically no longer than 10–15 tokens which are read by a range of different speakers. Readings longer than 10 seconds are discarded.

We followed the training process in Tyers and Meyer (2021) to train speech recognition models for Hungarian and Estonian using the Common Voice data. After training the models we extracted a number of keywords for the two languages from their test sets. We selected all tokens that appeared more than 5 times and that were 5 characters or longer. This second constraint was to try and avoid closed categories that would be unlikely to be used as keywords (e.g. Hungarian *és* 'and' or Estonian *on* 'is').

4.2 Real-word data

As the experiments with Common Voice demonstrate what can be done with read speech, we wanted to see how well the models would work with spontaneous speech of the type more typically found in language archives.

4.2.1 Finnish

The Finnish test data is taken from a CC-BY licensed Samples of Spoken Finnish corpus (Institute for the Languages of Finland, 2014), which contains 100 recordings of 50 Finnish dialects recorded primarily in the 1960s and 1970s. What makes this material particularly relevant is that the recordings originated in the Finnish dialect documentation program, which aimed to record 30 hours of dialect materials from each Finnish municipality. By the end of the 1970s the collections already contained 15,000 hours, and the currently available Finnish dialect materials, in the Institute for the Languages of Finland alone, number 24,000 hours[1]. The materials from which our sample is taken represents a tiny fragment of the recordings that have ever been published in any format.

We have selected five recordings from different dialect regions, and tagged the transcriptions for 100 keywords. The recordings chosen from the corpus were SKN03b_Palkane, SKN10b_Mikkeli, SKN12a_Salla, SKN13b_Pihtipudas and SKN18b_Rautalampi. The keyword tagging is applied on this dataset, and the accuracy is measured. We believe the Finnish results will be generalizable to the wider context of archived Finnish multimedia, at least what it comes to this portion of the dialect recordings. We used the normalized versions of the transcriptions, as those are available in the corpus we used. Those deviate in various ways from the original dialectal representation, but the high variation between word forms in different dialects would had made the comparison of keywords challenging. In the further work, the dialectal variants of the wordforms could be mapped together to allow more dialect-aware keyword search. At the same time, to our knowledge, no ASR system has yet been trained that would even start to address the phenomena met in the dialectal Finnish, and the target of these systems is usually modern literary Finnish. Also the current training data for our Finnish ASR model

[1] https://www.kotus.fi/aineistot/puhutun_kielen_aineistot

Source	Language	Autonym	Locale	Training	# Clips	# Speakers	$\|V\|$
Ardila et al. (2020)	Finnish	Suomi	`fi`	0:32:29	456	1	28
Ardila et al. (2020)	Hungarian	Magyar nyelv	`hu`	4:17:04	3339	2	36
Ardila et al. (2020)	Estonian	Eesti keel	`et`	5:00:16	2760	73	34
Hjortnaes et al. (2020)	Komi	Коми кыв	`kpv`	38:56:02	53711	232	60

Table 1: **Languages and data**. The datasets used in training the speech recognition models that were used in these experiments.

is basically in modern literary Finnish, as it was trained using the read sentences from Common Voice, making it poorly suited for dialectal data.

4.2.2 Komi

For Komi we used a story recorded by Erik Vàszolyi (for various versions of 'Ballad of the soft-haired sister' see Vászolyi-Vasse, 2001; Vászolyi-Vasse and Lázár, 2010), described in a recent study by (Blokland et al., 2021). This is a text that exists in two variants, as it has been recorded both as a sung and narrated version. The narrative version used in this experiment is 17 minutes long. This text is particularly relevant for testing keyword recognition, as it has culturally very relevant content to detect. However, the sang version of the text was already included in the training material of the model, invalidating any results obtained from testing on that data, and thereby excluded from comparison. Especially with the archival data, the same individual is often recorded numerous times, so a situation where some of their recordings are already included into the model is not entirely unrealistic. As always, further testing is obviously required with more speakers and text types. Also for Komi we manually selected 100 keywords that are represented in the text.

As this Komi text was recorded with a tape recorded in 1966, it is very representative of archived Komi materials that do exist in large quantities in different archives. We described the wider context of the archival recordings most familiar to us in Section 2. This illustrates how one central goal in work described here is to be able to better navigate and access untranscribed archival recordings. We describe the related methodology next.

5 Methodology

Keyword spotting is the task of finding specific words in a given audio stream, often containing continuous speech. This has a wide variety of uses, most notably *keyword search* and *wake-word detection*. Keyword searching is when you have a large collection of audio saved on disk, and you want to identify all the instances of certain word. This is especially useful for information retrieval scenarios, and is easily generalizable to the situations where we know something about the recordings, but not exactly where which topic is discussed.

The task discussed in this study, keyword spotting, is just one part of a larger pipeline that related technologies create. This involves text recognition of already written transcriptions, and forced alignment of the text with audio. Keyword spotting usually predates a well functioning ASR, as it can be, arguably, implemented before speech recognition is yet fully established. In the longer perspective keyword tagging is also related to subject indexing, where the topics and keywords are extracted from the document text. Such systems are already successfully in use with larger Uralic languages, such as Finnish (Suominen, 2019). Indeed, keyword spotting would regularly be conducted in a context where we have reasons to assume specific term of interest is used somewhere in the document, be that a text or recording.

While there are specific algorithms for keyword spotting, cf. Mazumder et al. (2021), we use a very simple approach. We decode the audio as if we are performing a normal Speech-to-Text transcription task, and then we do a simple text search over the transcript. In this study we did not use specific keyword boosting techniques, which would be an additional approach to improve the findability of a specific string. Such use cases also distinguish keyword spotting more clearly from speech recognition, as our current methodology essentially uses generated transcription as a starting point.

For the experiments, we took the test set for each language, and selected 10 words at random from a set of those words longer than four characters to favour content words over function words. The results are presented in Table 2.

Language	# Keywords	F_1	Prec	Rec
fi	100	0.15	0.41	0.09
hu	192	0.28	0.89	0.16
et	546	0.33	0.88	0.21
kpv	100	0.20	0.76	0.12

Table 2: **Keyword spotting**. We show the dataset size, precision, recall and F_1 score. In general the precision is high and recall is moderate to low.

6 Results

We will first explain the concepts we have used to measure the model's performance. Precision (Prec) is how often the model is correct when it identifies a keyword. Recall (Rec) is how many of the keywords in the test data the search is able to find. F_1 is a weighted average of precision and recall which tends towards whichever value is lower, meaning the best score is achieved by balancing precision and recall. This gives intuitively interpretable and comparative information about the experiments.

Our results were the best for Estonian and Hungarian. We believe this is largely connected to the narrow domain which was present in the Common Voice recordings, namely that the clips are read. The low accuracy of Finnish is probably related to the small amount of training data. Without an accurate model, the keywords may not be correctly transcribed and will not show up in the text search. In the case of Komi we reach a relatively high precision, on par with Hungarian and Estonian where the domain was narrow, and here the large amount of training data must have some role. However, the clips are from natural speech instead of read, which explains the lower accuracy when compared to Hungarian and Estonian despite the large quantity of training data. This is not an excellent result, but already a step toward a clearly functional system. As the recall is very low, it must stated that the system is not very successful in finding the keywords, but when it suggests them, those are often correct.

We expected Estonian and Hungarian to work relatively well, since the test data was not very realistic. However, the result with Komi comes relatively close to what we see with the test languages. Especially with Finnish experiments with more training data, possibly varying the training data size gradually, could help to understand how the ratio of the training data impacts to the model's performance. Similar experiment was previously conducted successfully for Kamas to evaluate changes in the accuracy (Partanen et al., 2020). We also have to emphasise that the Finnish data was much more strongly dialectal than what would be customarily encountered in the recordings today, and what is present in the Common Voice dataset. Even though such older dialect recordings exist in large quantities in Finnish archives, they must still be considered a special case within Finnish speech technologies in general.

Another challenge, and factor that makes our results less reliable, is that we selected the keywords from the corpora themselves. This was the only available approach, as we wanted to measure the accuracy, but it also targeted our experiment toward the existing inflected forms that do exist in the test data. With agglutinative Uralic languages, however, the most useful test scenario would be one where the desired keywords are listed by their lemmas, but may occur in a different shape in the real usage, and the keyword spotting would ideally still work.

7 Concluding remarks

Our research shows that keyword detection systems are in principle applicable for low resource settings, and even with a very small amount of training data the precision can be relatively high. It certainly is not possible to retrieve all keywords reliably under the current conditions, but even the accuracy we are now reaching could still be useful. Naturally, lots of work still remains to be done within this topic.

One of the most important further tasks would be to extend the experiment into entirely realistic conditions. We could, for example, use archived recordings and their keyword lists and summaries to create the keyword queries, and compare the result against manually verified data. This way we could move toward concrete evaluation of how well and realistically the system performs with various archived datasets. Also different fieldwork collections in Uralic languages could be very well suited for this task. Even though exact keyword and topic listings may not be very common in current metadata models, there is still a long tradition of compiling such topic indexes, and this is inarguably a very useful strategy to classify non-transcribed recordings. Combined to keyword spotting such index can be used to navigate the recordings as well. Our current study is a first step to that direction in a wider context of Uralic languages, and with the goal of trying to test the keyword detection in languages representing different branches of this language family.

Acknowledgments

Niko Partanen has produced this work within the project *Language Documentation meets Language Technology: The Next Step in the Description of Komi*, funded by Kone Foundation, Finland. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, et al. 2020. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. *arXiv preprint arXiv:2101.03027*.

Tanel Alumäe, Ottokar Tilk, et al. 2019. Advanced rich transcription system for Estonian speech. *arXiv preprint arXiv:1901.03601*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Rogier Blokland, Niko Partanen, and Michael Rießler. 2021. *This is thy brother's voice*. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual facilitation*. University of Helsinki.

Rogier Blokland, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA*, volume 2, pages 24–30.

Seppo Enarvi, Peter Smit, Sami Virpioja, and Mikko Kurimo. 2017. Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2085–2097.

Liis Ermus, Mari-Liis Kalvik, and Tiina Laansalu. 2019. The Archive of Estonian dialects and Finno-Ugric languages at the Institute of the Estonian language. *Uralica Helsingiensia*, (14):351–366.

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020. Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In Tommi A. Pirinen, Francis M. Tyers, and Michael Rießler, editors, *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37. Association for Computational Linguistics.

Nils Hjortnæs, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2021. The relevance of the source language in transfer learning for ASR. In Miikka Silfverberg, editor, *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages*, volume 1, pages 63–69. University of Colorado Boulder.

Institute for the Languages of Finland. 2014. Suomen kielen näytteitä - Samples of Spoken Finnish [online-corpus], version 1.0. http://urn.fi/urn:nbn:fi:lb-201407141.

Erkki Itkonen. 1958. Komin tasavallan kielitieteeseen tutustumassa. *Virittäjä*, 62(1):66–66.

Abhilash Jain, Aku Rouhe, Stig-Arne Grönroos, and Mikko Kurimo. 2020. Finnish asr with deep transformer models. In *Conference of the International Speech Communication Association (INTERSPEECH)*, volume 21.

Mark Mazumder, Colby Banbury, Josh Meyer, Pete Warden, and Vijay Janapa Reddi. 2021. Few-shot keyword spotting in any language. *arXiv preprint arXiv:2104.01454*.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech recognition for endangered and extinct samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Günter Johannes Stipa. 1962. Käynti syrjäänien tieteen tyyssijassa. *Virittäjä*, 66(1):61–68.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing ASR pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Osma Suominen. 2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, 1(29):1–25.

Ville T Turunen and Mikko Kurimo. 2008. Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(1):1–25.

Francis M. Tyers and Josh Meyer. 2021. What shall we do with an hour of data? speech recognition for the un- and under-served languages of common voice.

Muusa Vahros-Pertamo. 1963. Syrjäänien asuinseuduilla. *Virittäjä*, 67(1):77–85.

E. Vászolyi-Vasse. 1999. *Syrjaenica*, volume One of *Specimina Sibirica*. Seminar für Uralische Philologie der Berzsenyi Hochschule.

Erik Vászolyi-Vasse. 2001. *Syrjaenica*, volume 2. Seminar für Uralische Philologie der Berzsenyi Hochschule.

Erik Vászolyi-Vasse and Katalin Lázár. 2010. *Songs from Komiland*. Reguly Társaság.

Ewald van der Westhuizen, Herman Kamper, Raghav Menon, John Quinn, and Thomas Niesler. 2021. Feature learning for efficient ASR-free keyword spotting in low-resource languages. *Computer Speech & Language*, page 101275.

Guillaume Wisniewski, Alexis Michaud, Benjamin Galliot, Laurent Besacier, Séverine Guillaume, Katya Aplonova, and Guillaume Jacques. 2020. Ouvrir aux linguistes «de terrain» un accès à la transcription automatique. *Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, page 82.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Виктор Николаевич Денисов. 2014. Из истории первых фонографических записей удмуртов и коми-пермяков в 1911-1912 гг. На территории верхнего Прикамья. *Ежегодник финно-угорских исследований*, (4).