INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Implementation of a Radiology Speech Recognition System for Estonian using Open Source Software

Tanel Alumäe[1], *Andrus Paats*[1,2], *Ivo Fridolin*[1], *Einar Meister*[1]

[1]School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia
[2]North Estonian Medical Centre, Tallinn, Estonia

tanel.alumae@ttu.ee, andrus.paats@regionaalhaigla.ee, ivo@cb.ttu.ee, einar@ioc.ee

Abstract

Speech recognition has become increasingly popular in radiology reporting in the last decade. However, developing a speech recognition system for a new language in a highly specific domain requires a lot of resources, expert knowledge and skills. Therefore, commercial vendors do not offer ready-made radiology speech recognition systems for less-resourced languages.

This paper describes the implementation of a radiology speech recognition system for Estonian, a language with less than one million native speakers. The system was developed in partnership with a hospital that provided a corpus of written reports for language modeling purposes. Rewrite rules for pre-processing training texts and postprocessing recognition results were created manually based on a small parallel corpus created by the hospital's radiologists, using the Thrax toolkit. Deep neural network based acoustic models were trained based on 216 hours of out-of-domain data and adapted on 14 hours of spoken radiology data, using the Kaldi toolkit. The current word error rate of the system is 5.4%. The system is in active use in real clinical environment.

Index Terms: speech recognition, radiology, medical dictation, open source, less-resourced languages

1. Introduction

Radiological reporting is a field in which automatic speech recognition (ASR) technology has been very successful. The first application of isolated word speech recognition for radiology was described already in 1981 [1]. More widespread use of ASR in radiology started more than a decade later when continuous speech recognition became more mature [2]. Many large-scale deployments demonstrated that integration of ASR into radiology decreased departmental operational costs and radiology report turnaround times [3, 4, 5, 6].

Although there have been many studies analyzing the impact, advantages and productivity issues of integrating ASR into radiology reporting workflow, the technical details of the actual radiology ASR systems have received little attention in academic publications. In fact, we were able to find only one paper [7] that describes all the components and technical considerations behind a radiology ASR system.

Estonian is the official language of Estonia, spoken natively by about one million people. Commercial speech recognition support for Estonian is non-existent. Even in the field of radiology, there aren't any commercial offerings for Estonian available.

This paper gives a detailed description of the Estonian radiology ASR system developed by Tallinn University of Technology and North-Estonian Medical Centre (NEMC). The first prototype of the system, using Gaussian Mixture Model (GMM) based acoustic model (AM) was described in [8]. Since then, many improvements have been made to the system, including the integration of domain-adapted deep neural network (DNN) based AMs, language model (LM) adaptation using real dictated texts, smarter handling of sentence breaks and spoken noises in the language model, that have together reduced the word error rate (WER) to around 5%, an almost 75% relative reduction compared to the first prototype. The system uses only free and open-source technology, most notably Kaldi [9] for ASR and Thrax [10] for implementing context-sensitive rewrite rules that convert text to spoken form and back. The system is now daily used by radiologists.

The paper serves as a case study on implementing an accurate domain-specific ASR system using relatively little resources and only open source software. The acoustic models were trained using freely available 216 hours of speech from very different domains and adapted using only 14 hours of in-domain data. We acknowledge that this does not really qualify as an under-resourced setting. However, the increase in the amount of available training data for Estonian in the last 10 years has been due to the corpus collection work at Tallinn University of Technology [11], funded by the National Program for Estonian Language Technology[1] which serves as a good example on how to turn a less-resourced language into a well-resources language using systematic work. The paper also shows how different steps in improving acoustic and language models affect the final system performance.

2. System description

2.1. Background

Estonian language belongs to the Finno-Ugric language family. It is closely related to Finnish and more distantly to Hungarian. Estonian employs the Latin script as the basis for its alphabet. It is a language with a close relationship between word orthography and pronunciation. Estonian is a highly infectional language: nouns and adjectives decline in fourteen cases. It is also a heavily compounding language: new lexemes can be formed spontaneously from two or more shorter words to denote new concepts (i.e., as in the the English word *footpath*).

In many countries, the traditional radiology reporting workflow involves a radiologist who dictates the report to a tape recorder, and an assistant who manually transcribes the dictated report. In Estonia, however, radiologists traditionally enter the report by computer keyboard into the radiology information system (RIS) themselves. As a result, the typed reports tend to be quite heavy in abbreviations and acronyms, contain relatively many spelling errors and use inconsistent formatting (e.g., some radiologists always use all-uppercase letters in subheadings).

[1] http://www.keeletehnoloogia.ee

http://dx.doi.org/10.21437/Interspeech.2017-928

Table 1: *Sample written expressions and their corresponding verbalized forms.*

Written	Verbalized
see , on test .	see koma on test punkt
d. cysticus'es	duktus tsüstikuses
D35 8F L4-S1	D kolmkümmend viis kaheksa F L neli S üks
A/B-uuring	A B uuring
45 x 4,6 cm tsüst	nelikümmend viis korda neli koma kuus sentimeetrine tsüst
teisipäeval 30.04. kl 10:25 .	teisipäeval kolmekümnendal aprillil kell kümme kakskümmend viis punkt

2.2. Data

We used 642 033 written radiology reports produced between 2010 and 2014 at NEMC for language modeling purposes.

In order to develop and verify context-sensitive rewrite rules for non-standard words, we asked two radiologists to manually verbalize 1000 randomly selected reports. Verbalization refers in this paper to the task of converting text from a written representation into a representation of how the text is to be spoken, while still using orthographic notation. Verbalization included expanding numbers and abbreviations to fully spoken forms and Estonian phonetic transliteration of Latin words and foreign names. Table 1 lists some example written expressions and their verbalizations.

During the development phase, we performed two intermediate tests and a third final test, where we asked radiologists (18, 12 and 11 unique speakers, respectively) to use the current system prototype to dictate real radiology reports in real clinical environment. We collected the speech and manually verified the corresponding report texts. The number of reports in each of the test sets is 364, 516 and 219, respectively.

2.3. Acoustic modeling

Acoustic models were trained using the Kaldi toolkit [9]. We trained the DNN models using 216 hours out-of-domain speech data from various domains, mostly collected at Tallinn University of Technology in the last 10 years (see Table 2) [11]. The models were later adapted using 14 hours of in-domain data that we collected during the two intermediate system prototype tests.

The acoustic model inventory contains 43 phone models, a silence model that is also used for modelling various filler words, and a garbage model used to absorb unintelligible and foreign language words during training.

The acoustic model training follows the Kaldi recipe for training time-delay DNN (TDNN) models [12]. First, a speaker-adaptive GMM system was trained, that was used for generating state-level alignments for training the TDNN acoustic model. The TDNN was trained on three speed-perturbed copies of the acoustic training data, with corresponding speed factors of 0.9, 1.0 and 1.1, and utterance-specific random volume perturbation using random scaling factors in the range of $[1/8..2]$ [13]. Unnormalized untruncated 40-dimensional MFCC features are used as input. The TDNN has five hidden layers that use the p-norm non-linearity [14], each with an input dimensionality of 3500 and output dimensionality of 350. The splicing indices per layer are `-2,-1,0,1,2 -1,2 0 -3,3 -7,2`, resulting in an effective left context width of 13 and right context width of 9. 100-dimensional i-vectors extracted in online manner are used as additional inputs to the TDNN. TDNN targets correspond to the 6770 context-dependent tied phone states. The model was first trained over three epochs with the cross-entropy (CE) criterion, using the out-of-domain data. Then, it was finetuned on speed-perturbed in-domain data, using three epochs of cross-entropy training and five epochs of sequence-discriminative training. We also tried other approaches to model adaptation using in-domain data, including adding a new hidden layer to the baseline acoustic model, using KL-divergence based regularization when finetuning the model [15], retraining a new model on in-domain data, using hard alignments interpolated with soft targets generated by the background model, but none of those approaches worked better than the method described above.

Table 2: *Out-of-domain acoustic model training data.*

Data	Amount (hours)
Broadcast conversations	109
Lectures and conference speeches	38
Broadcast news	30
Studio-recorded spontaneous speech	29
Phonetically balanced dictated speech	8
Data from our Android voice input app	2
Total	*216*

A rule based system is used for deriving the pronunciations for words in the LM lexicon [2]. For many common foreign proper names and abbreviations, pronunciation is created by first transforming the lexical form to a localized form using a transliteration table, and then applying the common pronunciation rules.

2.4. Language modeling

A radiology ASR system has to decode verbalized language and convert it into properly normalized and formatted written language for presentation to the user. This is challenging for several reasons. The LM training data comes in written form, and determining the verbalized form of many non-standard words is not trivial. It is especially challenging in highly inflective languages like Estonian, where the verbalization of numbers and abbrebiations depends on the syntactic context. For example, an abbreviation "*cm*" is verbalized as "*sentimeeter*", "*sentimeetri*", "*sentimeetrit*" or "*sentimeetriga*" (among others), depending on the context. In the radiology domain, determining the verbalized form of many non-standard words requires domain knowledge. Converting the decoded verbalized text back into written form is also often ambiguous. In fact, creating the verbalization and normalization rules was one of the most time-consuming parts of developing the system.

2.4.1. Text verbalization and deverbalization

We adopted the finite state transducer (FST) based approach to LM verbalization and text normalization. Our method is somewhat similar to the approach described in [16], with several important differences. We also represent verbalizer rewrite rules as a finite state transducer, and use the inverse of this transducer to convert the decoded verbalized text back into written form. However, we don't compose the verbalizer model with

[2] Available at `http://github.com/alumae/et-g2p`

the decoding graph (HCLG) but apply the verbalizer offline to pre-process LM training text and apply the inverse of the verbalizer in online mode to post-process the decoded output. This allows us to use context-sensitive rewriting for ambiguous rules, contrary to the approach proposed in [16].

Our verbalizer model is a FST that is compiled from Thrax [10] grammar. The grammar contains rules for expanding numbers and dates into words, transliterating Latin words, expanding abbreviations and punctuation marks, expanding acronyms into letter-by-letter sequences. The rules were created manually in a test-driven manner by trying to minimize the difference between the manually verbalized reports created by radiologists and the automatically verbalized versions of the same reports. Later, the rules were finetuned based on the two intermediate tests.

The rules are mostly context-independent and often highly ambiguous. For example, the words "*10 cm*" are verbalized into more than 50 different strings, corresponding to different inflections of the words "kümme" ("ten") and "sentimeeter" ("centimeter"), among others:

- *kümne sentimeetrise*
- *kümnele sentimeetri**
- *kümne sentimeeter**
- *kümme sentimeetrit*

Note that that the second and third expansions are actually syntactically incorrect, since the inflection of the number "10" doesn't agree with the inflection of the following unit. This happens because the number and the unit are expanded using different rules and we don't attempt to make them agree, as it would be very complicated. In other words, our verbalizer model over-generates.

In order to inject context-sensitiveness into the verbalizer model, we compose it with an n-gram LM, encoded as a weighted FST. The n-gram is estimated on the manually verbalized corpus of 1000 radiology reports. This is similar to the method proposed for expanding Russian numbers to words for speech synthesis [17]. However, since the amount of training data is very small, we dont't pick the verbalization with the highest probability when pre-processing training data, but instead sample from the different verbalizations, with probability that is proportional to the posterior probability of the individual sentence verbalizations. This ensures that a wide variety of different verbalizations are actually represented in the training data. For de-verbalization, we invert the verbalizer model and compose it with a n-gram FST estimated from written reports. In this case, there is abundant training data for the n-gram model and we always pick the de-verbalization with the highest probability.

Using sampling instead of the best-path for verbalizing the LM training data results in the fact that some of the training data verbalizations are syntactically invalid, but in practice it should have little negative effect on the quality of the LM. Even if this causes the system to make small recognition errors (e.g., recognize "sentimeetri" instead of "sentimteerit"), such errors are usually hidden during deverbalization (both words in the previous example are transformed back to "cm").

2.4.2. Language model training

Before the LM estimation, the training texts were pre-processed as follows: the texts were tokenized and converted into verbalized form, using the model described in the previous section. We also replaced all newline characters with a special *<newline>* token, which was mapped to the pronunciation of the words "new line" in the lexicon. We also added random line breaks between any two words with the prior probability of 5%. This was needed because we observed that radiologists don't dictate the reports sentence-by-sentence but often make longer pauses in the middle of sentences, and on the other hand, dictate many simpler sentences without pauses between them. As a longer pause causes our system to finish the current utterance decoding and start a new one, the LM needs a capability to model intra-sentence n-grams as if they would be utterance starting. We also randomly insert a special word mapped to a garbage model between words with a probability of 5%, as this helps to model cases when non-stationary noises appear during dictation pauses. The garbage word is converted to a blank word during deverbalization.

Table 3: *Perplexities of various language models on development data.*

Model	Training data	PPL
KN-smoothed 4-gram	Written (1 year)	60
KN-smoothed 4-gram	Written (5 year)	49
KN-smoothed 4-gram	Written + Spoken	48
MaxEnt 4-gram	Written + Spoken	45
MaxEnt 4-gram, adapted to spoken data	Written + Spoken	42

The LM vocabulary is created by selecting all words from the pre-processed training texts that appear at least four times. This results in a vocabulary of 62 108 words that has an out-of-vocabulary (OOV) rate of 0.62% against the development data. In contrast, the OOV-rate of a 60 000 word vocabulary in Estonian broadcast news domain is typically more than 10% and we usually need to apply morphological decomposition in the LM in order to make large vocabulary ASR feasible [18]. The main causes for such vocabulary explosion are morphological variety (i.e., nouns, adjectives and verbs occurs in many different inflected forms) and frequent use of compound words. It seems that the more restricted nature of the radiology language avoids such vocabulary explosion and we don't need to use morphological decomposition in the LM for the radiology domain.

The LM is a 4-gram model trained with the maximum entropy (MaxEnt) criterion [19] using the SRILM toolkit [20]. The model was first trained on the union of written reports and manually transcribed spoken reports from our two intermediate system tests. It was then adapted using only the spoken report data, using the parameters of the baseline model as a priors [21]. Table 3 lists LM perplexities of a Kneser-Ney (KN) smoothed and MaxEnt n-gram models trained over union of the written and spoken data and the MaxEnt model adapted to spoken data.

2.5. Architecture

We use a scalable distributed client-server based speech-to-text architecture as the backend of the system [22]. The server component consists of a master and one or more worker pools [3]. The master server delegates recognition tasks from clients to workers that can be deployed on one or more remote computers. The system is confugured to perform decoding in almost real-time. The system client is very lightweight and was implemented as a Java application. A new Javascript-based client is being integrated into the hospital's RIS.

[3] Available at `http://github.com/alumae/kaldi-gstreamer-server`

Table 4: *Word error rates after different system development stages.*

ID	System description, relative to previous	WER
A	GMM AM trained on out-of-domain data, N-gram LM trained on 1 year of written reports	19.1
B	DNN AM trained on out-of-domain data	13.8
C	N-gram LM trained on 5 years of written reports	13.6
D	Insert a garbage word (recognized as blank) between any two words (with $p = 0.05$) in LM training data	8.9
E	Concatenate sentences in LM training data and break them randomly (with $p = 0.05$)	8.4
F	Longer silence threshold for utterance end detection	7.8
G	DNN adapted to in-domain data using CE training	5.7
H	DNN adapted to in-domain data using CE and sequence-discriminative training	5.5
I	MaxEnt LM adapted using spoken data	5.4

Table 5: *Word error rates of the final system on different radiology modalities. The second and third columns show the number of test reports and the average number of words per report.*

Modality	#	Words/report	WER
Computed Tomography	87	127	5.4
Ultrasound	47	75	4.0
Magnetic Resonance	42	90	6.7
X-ray	42	37	5.2
Other	1	74	6.8

3. Evaluation

The system was evaluated on the final test data which includes 219 dictated radiology reports by 11 different radiologists.

In order to assess the importance and effect of different system development steps, we ran simulations with the test data using earlier system prototype versions. We started with a GMM-based system that was described in [8] and ended with the final system prototype. Table 4 lists average WER results after each improvement. The speaker-specific and average WERs corresponding to the stages listed in Table 4 are plotted in Figure 1.

It is not surprising that the DNN AM gives a large 28% relative WER improvement over the GMM-based AM (A $\rightarrow$ B). The large 35% relative improvement brought by inserting a special unit represented by a garbage model (recognized as blank) between any two words in the LM training data (C $\rightarrow$ D) is related to the fact that many test speakers preferred to leave microphone open between individual utterances while investigating the radiology image. The background noises and mouse clicks during such silence regions caused the system to hallucinate short words into the recognized text stream, resulting in many insertion errors. Allocating some probability mass to a garbage model in the LM largely fixed this problem, although it probably could be also achieved by inserting a dedicated speech detection module in our recognition pipeline. Another large 27% relative improvement (F $\rightarrow$ G) comes from adapting AMs using in-domain data. Finally, it is perhaps surprising that LMs

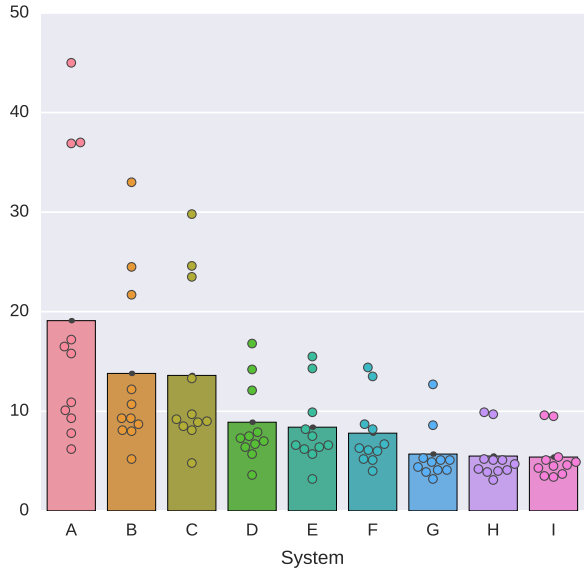

Figure 1: *Word error rates corresponding to individual radiologists in different system development stages (marked with dots) and the corresponding average WERs (marked with bars). System IDs correspond to those in Table 4.*

with a better perplexity result in relatively little WER improvements. We suspect this is because the baseline LM is already very accurate and most of the ASR errors are not caused by a weak LM.

Table 5 shows some test statistics about the different radiology modalities. We see that the reports dictated for complicated 3D modalities (CT, MR) have higher WER values than X-ray and ultrasound. The results confirm findings from an earlier study [23] and from our previous results [8], although the relative difference between the modalities is now smaller.

4. Conclusions

We described the technical aspects of building a speech recognition system for Estonian. The system is based on open-source software, mainly Kaldi and Thrax. During the development phases, we were able to reduce the 19.1% word error rate of the first prototype to 5.4%. The most critical steps in improving the performance of the system were migration to DNN acoustic models, better handling of non-stationary noises between speech segments and adapting acoustic models using in-domain speech data.

Although the system is already used daily in the hospital, there are some critical issues that need to be addressed in order to make the system sustainable in the longer run: system vocabulary should be updatable without the need of an ASR specialist, and users should be able to add new simple rewrite rules without learning the internals of the system.

5. Acknowledgements

This work was partly funded by the Estonian Ministry of Education and Research target-financed research theme no. 0140007s12 and through the project Estonian Speech Recognition System for Medical Applications.

6. References

[1] B. Leeming, D. Porter, J. Jackson, H. Bleich, and M. Simon, "Computerized radiologic reporting with voice data-entry." *Radiology*, vol. 138, no. 3, pp. 585–588, 1981.

[2] K. S. White, "Speech recognition implementation in radiology," *Pediatric Radiology*, vol. 35, no. 9, pp. 841–846, 2005.

[3] S. Sferrella, "Success with voice recognition." *Radiology Management*, vol. 25, no. 3, pp. 42–49, 2002.

[4] M. R. Ramaswamy, G. Chaljub, O. Esch, D. D. Fanning, and E. vanSonnenberg, "Continuous speech recognition in MR imaging reporting: advantages, disadvantages, and impact," *American Journal of Roentgenology*, vol. 174, no. 3, pp. 617–622, 2000.

[5] J. D. Houston and F. W. Rupp, "Experience with implementation of a radiology speech recognition system," *Journal of Digital Imaging*, vol. 13, no. 3, p. 124, 2000.

[6] P. J. Lemme and R. L. Morin, "The implementation of speech recognition in an electronic radiology practice," *Journal of Digital Imaging*, vol. 13, no. 1, pp. 153–154, 2000.

[7] B. Angelini, G. Antoniol, F. Brugnara, M. Cettolo, M. Federico, R. Fiutem, and G. Lazzari, "Radiological reporting by speech recognition: the A.Re.S. system." in *ICSLP*, 1994.

[8] A. Paats, T. Alumäe, E. Meister, and I. Fridolin, "Evaluation of automatic speech recognition prototype for Estonian language in radiology domain: A pilot study," in *16th Nordic-Baltic Conference on Biomedical Engineering*. Springer, 2015, pp. 96–99.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU Workshop*, 2011.

[10] T. Tai, W. Skut, and R. Sproat, "Thrax: An open source grammar compiler built on OpenFst," in *IEEE ASRU Workshop*, 2011.

[11] E. Meister, L. Meister, and R. Metsvahi, "New speech corpora at IoC," in *XXVII Fonetiikan päivät*, 2012, p. 30.

[12] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *Interspeech*, 2015, pp. 3214–3218.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *Interspeech*, 2015.

[14] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *ICASSP*, 2014, pp. 215–219.

[15] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, 2013, pp. 7893–7897.

[16] H. Sak, F. Beaufays, K. Nakajima, and C. Allauzen, "Language model verbalization for automatic speech recognition," in *ICASSP*, 2013, pp. 8262–8266.

[17] R. Sproat, "Lightly supervised learning of text normalization: Russian number names," in *IEEE SLT Workshop*, 2010, pp. 436–441.

[18] T. Alumäe, "Recent improvements in Estonian LVCSR." in *SLTU*, 2014.

[19] T. Alumäe and M. Kurimo, "Efficient estimation of maximum entropy language models with n-gram features: An SRILM extension," in *Interspeech*, 2010.

[20] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," in *IEEE ASRU Workshop*, vol. 5, 2011.

[21] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *Computer Speech & Language*, vol. 20, no. 4, pp. 382–399, 2006.

[22] T. Alumäe, "Full-duplex speech-to-text system for Estonian," in *Baltic HLT*, 2014.

[23] S. Basma, B. Lord, L. M. Jacks, M. Rizk, and A. M. Scaranelo, "Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription," *American Journal of Roentgenology*, vol. 197, no. 4, pp. 923–927, 2011.