Error Analysis and Improving Speech Recognition for Latvian language

Askars Salimbajevs
Tilde, Vienibas gatve 75a, Riga, Latvia
askars.salimbajevs@tilde.lv

Jevgenijs Strigins
Tilde, Vienibas gatve 75a, Riga, Latvia
jevgenijs.strigins@tilde.lv

Abstract

Developing a large vocabulary automatic speech recognition system is a very difficult task, due to the high variations in domain and acoustic variability. This task is even more difficult for the Latvian language, which is very rich morphologically and in which one word can have dozens of surface forms. Although there is some research on speech recognition for Latvian, Latvian ASR remains behind “big” languages such as English, German etc. In order to improve the performance of Latvian ASR, it is important to understand what errors does it make and why. In this paper, the authors analyze the most common errors of Latvian ASR. Based on this, baseline system WER is improved from 30.94% to 28.43%.

1 Introduction

When developing an Automatic Speech Recognition (ASR) system it is typical to evaluate system performance by calculating quantitative measures like accuracy, F1 score, Word Error Rate (WER) etc. However, in order to improve ASR performance, it is important to understand which factors are most problematic for recognition, identify the types of errors, their main causes and how critical these errors are. This is even more important when developing ASR for a language, for which no such analysis has ever been done, because the developer might not know what problems to expect, where more effort and focus is needed and what possible solutions there are.

The Latvian language is a moderately inflected language, with complex nominal and verbal morphology. Latvian also has a selection of prefixes and suffixes that can modify nouns, adjectives, adverbs and verbs. There is no definite or indefinite article in Latvian, but definiteness can be indicated by the endings of adjectives. Because of these properties, one word in Latvian can have tens or even hundreds (in the case of verbs) of surface forms. For example, a word “cat” in English has 3 surface forms: *cat*, *cats* and *cat’s*, but in Latvian the variation is much bigger: *kaķis, kaķa, kaķim, kaķi, kaķī, kaķu, kaķiem* etc. They all describe an animal – cat, but in the same time these are different surface forms that change the meaning of sentence.

To the best of our knowledge there has been no research conducted on analyzing misrecognized words for Latvian LVASR. In fact, there are only a few published results on speech recognition for Latvian (Oparin et al., 2013; Darģis, R., & Znotiņš, A., 2014; Salimbajevs & Pinnis, 2014), that report the best performance of WER 20.2%.

However, there are no lack of efforts on error analysis for “bigger” languages (Goldwater et al., 2010; Vasilescu et al, 2012). In many cases factors discovered in these works also apply to “smaller” languages. For example there are results for English (Fosler-Lussier & Morgan, 1999) and Japanese (Shinozaki & Furui, 2001) that show that infrequent words are more likely to be misrecognized, which is most likely to be true also for other languages.

Most studies analyze errors from the perspective of the ASR vs. human capacities in decoding spoken signals and consider ASR errors from lexical or phonetic standpoints. There are, however, also efforts that focus on morpho-syntactic structure (Goryainova, 2014).

In this paper we present an error analysis of the Latvian Large Vocabulary Automatic Speech Recognition (LVASR) system. We do not perform in-depth analysis of ASR error causes, but rather concentrate on typical surface errors to

Proceedings of Recent Advances in Natural Language Processing, pages 563–569,
Hissar, Bulgaria, Sep 7–9 2015.

produce classes of errors and find general solutions.

The remainder of the paper is organized as follows. Section 2 describes the present Latvian ASR system used in this study. Classes of errors, their effect on utterance meaning and their causes are discussed in Section 3. Section 4 describes improvements which we have made after analyzing errors and gives a short evaluation of the improved system. All results of this study are then interpreted in Section 5. Section 6 concludes the paper.

2 Latvian ASR

The present Latvian Automatic Speech Recognition system is based on an open-source Kaldi toolkit (Povey et al., 2011), which in turn is based on the Weighted Finite State Transducer (WSFT) approach. We use this system as a baseline for analyzing recognition errors and testing improvement ideas. The system's details are described in the following subsections.

A few results on Latvia

2.1 Acoustic Modelling

The acoustic model (AM) is trained on a 100 hour-long Latvian Speech Recognition Corpus (Pinnis et al., 2011). We use the following acoustic model setup:

- HMM (hidden Markov models)-DNN (deep neural network) modelling approach.
- MFCC features and LDA. These are 40-dimensional feature vectors that are calculated from audio signal, and are used in actual calculations
- 37 base phonemes.
- 1 unified filler\silence model. Fillers represent sounds that are not spoken words, such as breathing, laughing etc.
- 1 garbage model for fragmented words and other garbage. For example, if word was not fully pronounced.
- iVectors are used for speaker adaptation (Miao et al., 2014). That is, for each speaker model parameters are changed so the better fit is obtained.

2.2 Language Modelling

The baseline ASR system uses n-gram language models (LM) which are trained on a 22M sentence and 304M word text corpus, which was collected by crawling Latvian web news portals. A vocabulary of 200K units is used, selected by their frequency in the training corpus.

Two language models are used during recognition:

- A 2-gram heavily pruned model is used during first-pass.
- A full not-pruned 3-gram model is used for rescoring lattices.

3 Recognition Errors

Here we used a small (approximately 23 minutes) corpus of Latvian speech, which was obtained by recording various people reading internet web news. The corpus was divided into two equal parts:

- A development set which is used for error analysis and testing possible improvements.
- A test set which is used to evaluate an improved speech recognition system.

Division was performed by randomly dividing this Latvian speech corpus in two parts with approximately equal length and same speakers.

3.1 Types of Errors

First we classified all errors by the following criteria:

- Whether the error is in the ending of the word.
- Whether the error is in a short word (we classify a word as short if it is no longer than 3 letters)
- Whether word boundaries were misaligned e.g. when the second part of one word is recognized as a part of the next word.
- Whether the previous word was recognized incorrectly.
- Whether the correct word is substituted with other word(s).

Using this criteria ASR output was compared with the correct transcripts. A summary of analyzed data is presented in the table below:

Category	% of All Errors
Ending	41%
Short word	15%
Word boundaries	13%
Error in previous word	28%
Substitution	52%

Table 1: Error summary from analyzing transcripts.

The table shows the percentage of specific categories of errors from all errors. It is important to analyze these categories, because, endings define different inflections. Short words are hard to discriminate acoustically and often they are partially skipped or spelled incompletely during fast human speech. Incorrectly defined word boundaries and word substitutions are common errors for speech recognizers. If one word is incorrectly recognized, then wrong n-grams of language model will be used in the process of calculating the probability of word sequence, therefore an effect of wrongly recognized previous word must be measured.

As our error categories overlap the total can be more than 100%. It can be seen that because of the inflective nature of Latvian, a large amount of misrecognitions are incorrect surface forms. For example, if the correct word is *kaķis*, but recognition output is *kaķi*, then it will be treated as an error, although these are actually different inflections representing the same word.

The usual output of a speech recognition decoder is a lattice of words, containing their estimated acoustic and language model costs. We used lattices to look deeper and classify errors using the following criteria:

- Whether words preceding or succeeding the wrongly recognized word are out of vocabulary words.
- Whether the correct word is in the lattice i.e. corrected word was pruned and cannot be recovered by rescoring.
- Whether the AM cost is too high (the incorrect word or surface form has a lower cost).
- Whether the LM cost is too high.

This means that in the case of an incorrectly recognized word, we investigate whether the correct word was actually present in the lattice along the best path, and if it was we compare the acoustic and language model scores of correct and recognized words. A summary of the lattice analysis is presented in Table 2.

Category	% of All Errors
Pruned from lattice	45%
Bad AM score	67%
Bad LM score	51%

Table 2: Error summary from analyzing lattices.

It can be seen that 45% of misrecognized words were pruned from lattices. Table 2 also suggests that there are more errors with bad AM cost, but this data is not sufficient to make any conclusions.

While performing this analysis we found that none of the fractional numbers were recognized correctly because of misrecognition of the word “komats” (decimal comma). We will investigate the cause of this problem further in the next paragraphs.

3.2 Effect of Errors

Not all recognition errors are equally important. For example a user will most likely be able to understand a transcript with errors in word endings, but completely misrecognized words (especially OOV words) and numbers can significantly change the meaning of utterances. These words can carry such critical data as time, person names, places etc. There have been attempts to automatically detect errors in critical words and use different clarification strategies to resolve them (Stoyanchev et al., 2012; Pappu et al., 2014).

First we analyzed the error distribution between parts of speech (POS) and how many of these errors are in word endings.

The second column in Table 3 shows what percentage of each part of speech is not recognized correctly. It can be seen that adjectives, verbs, particles and prepositions are the most difficult to recognize. Misrecognized verbs are more critical, as they can change the meaning of utterance or make whole utterance meaningless. Misrecognized adjectives are less important, as 75% of these misrecognitions are errors in endings, which should not make utterance unintelligible. Particles and prepositions are also less important for recovering the meaning of utterance.

Although there is no inflections for particles and prepositions, these are hard to recognize because usually they are short words that are not spelled very clearly during human speech and can become part of other words.

POS	% of Misrecognitions	% of Ending Errors
Adjectives	20%	75%
Conjunctions	12%	-
Nouns	16%	34%
Numerals	13%	53%
Particles	20%	50%
Participles	12%	57%
Prepositions	20%	-
Verbs	20%	45%
Other	10%	51%

Table 3: Errors in Parts of Speech.

Next we performed a subjective evaluation of recognized utterances. In total, 56% of utterances contain one or more errors that make it very difficult or impossible to recover the original meaning, while 47% of errors were critical. This result shows that the usability of transcriptions made with the current ASR can be very limited if no audio is available to check suspicious or important places in the text.

Also, while analyzing utterances we confirmed that OOV errors are critical for recovering the meaning of utterances. 82% of OOV errors significantly changed the meaning of utterances.

3.3 Causes of Errors

Analysis of the transcripts and lattices led to a number of hypotheses about the causes of different types of errors. In this section we list and test these hypotheses.

3.3.1 Word Length

Of particular interest was whether short words are harder for ASR to recognize than long ones. Let us define the probability of ASR wrongly recognizing short and long words by *P(s)* and *P(l)* respectively. A maximum likelihood estimate for these probabilities would be

$$\tilde{P}(s) = \frac{cnt(e_s)}{cnt(sw)}; \tilde{P}(l) = \frac{cnt(e_l)}{cnt(lw)}$$

where $cnt(e_s)$ and $cnt(e_l)$ are counts of errors in short words and long words, but $cnt(sw)$ and $cnt(lw)$ are the total count of short and long words in the corpus. The estimates are compared using Welch's t test to test the following hypotheses:

$$H_0: \tilde{P}(s) = \tilde{P}(l)$$
$$H_1: \tilde{P}(s) < \tilde{P}(l)$$

The statistical test yields a p-value of 0.001956, which is strong evidence against the null hypothesis. This result appears to be quite confusing, as short words are easier for ASR to recognize than longer ones.

3.3.2 Misrecognized Previous Word

Another issue is the effect of a wrongly recognized word on the recognition of the next word. Let us define the probability of ASR wrongly recognizing the current word given that the previous word was wrongly recognized by $P_{-1}(e)$ and the probability of ASR wrongly recognizing the current word given that the previous word was recognized correctly by $P_{-1}(c)$. The maximum likelihood estimates of these probabilities would be

$$\tilde{P}_{-1}(e) = \frac{cnt(w_e w_e)}{cnt(w_e)}; \tilde{P}_{-1}(c) = \frac{cnt(w_c w_e)}{cnt(w_c)}$$

Where $cnt(w_e w_e)$ is a count of the sequences of two consecutive errors, $cnt(w_c w_e)$ is a count of the sequences of an error preceeded by a correctly recognized word and $cnt(w_e)$ $cnt(w_c)$ would be the total number of errors and correct words. Then we would test the following hypotheses:

$$H_0: \tilde{P}_{-1}(e) = \tilde{P}_{-1}(c)$$
$$H_1: \tilde{P}_{-1}(e) > \tilde{P}_{-1}(c)$$

This statistical test yields a p-value of 0.8e-6, which is strong evidence against the null hypothesis. The estimated probabilities are 28% and 12%, which means that the previously incorrectly recognized word increases the probability of recognizing the next word incorrectly by more than 2 times.

3.3.3 Weak Decoding LM

As we have already seen, the correct words were pruned from the lattices in 45% of cases. Our hypothesis was that 2-gram pruned LM used in decoding would assign the wrong costs.

To test this hypothesis we made several experiments where a bigger 3-gram LM was used in decoding. We also tried to increase the lattice beam so that fewer paths are pruned. However, despite a decrease in the percentage of pruned words, no improvement was observed.

We also made a short analysis of cases where the correct word was still present in the lattice, but had a worse LM or AM cost (Table 4). The LM and AM costs are inversely proportional to probabilities of corresponding hypothesis obtained from *Kaldi* speech recognition toolkit.

Type	%
Only LM cost	30%
Only AM cost	46%
Both	24%

Table 4: Incorrect costs in lattices.

In a majority of cases the correct word was not chosen because it had a worse acoustic score and the LM cost was not small enough for correct variant (or large enough for the incorrect variant) to compensate for this. This result shows that our hypothesis was false and improvements in both AM and LM (both decoding and rescoring) are needed.

3.3.4 Out-of-Vocabulary

If the word is not in the system's vocabulary, it cannot be recognized. Moreover, an out-of-vocabulary word is known to generate between 1.5 and 2 errors (Schwartz et al, 1994). This is an important problem for Latvian ASR, because each word in Latvian has many surface forms and all of them must be in the vocabulary.

We found out that OOV words contribute to 13% of recognition errors. Also 5% of misrecognized words preceded or succeeded OOV words.

3.3.5 Misrecognition of Word “komats”

We identified two reasons for incorrect recognition of the word “komats” (comma). The first is pronunciation. Many people pronounce “komats” as “koma” which is a different word. The second is an excessively high LM cost for numerals and “komats”. LM is trained on a written text, but it is rare for numerals to be written using words, so numbers are written mostly using digits. Our baseline training procedure does not have any number to word conversion and all sentences with such numbers are filtered. Hence n-grams with numerals and “comma” are very rare, their probability is estimated as low and they can be pruned from decoding LM.

As a result both costs of word “komats” were high, so it was never chosen, instead some completely different words were chosen as the final hypothesis, making it very difficult to understand the meaning of the utterance.

4 Improving Latvian ASR

After analyzing error types, their importance and causes, the next step was to find ways to improve the current baseline system. In this section we describe our efforts to deal with some type of errors that we identified earlier.

We first tested individual improvement ideas on our development set. Then all the improvements were combined together and the improved system was evaluated on a test set.

4.1 Recognition of Word “komats”

The first step was adding the alternative pronunciation “komats = K O M A” in the grapheme-to-phoneme (G2P) dictionary. With this simple solution we achieved a 50% reduction of errors for the word “komats”.

The next step was implementing a number conversion step (done with a custom python script) in our LM training procedure. Implementing such a converter for Latvian is challenging, because all word endings must be matched. Our implementation covers only basic cases. After these efforts, 25% of the remaining errors with “komats” were corrected.

4.2 Word Endings

Table 1 shows that 41% of all errors are caused by misrecognized endings. Our solution to this problem involves increasing the language model training corpus from 22M to 47M sentences, while leaving the vocabulary size at 200K units. This should help to better estimate bigrams and trigrams, which contain words with rare endings, compared to the estimate obtained using backoff and a smaller corpus.

This approach led to a decrease of word ending errors to 21%, although there was no WER improvement.

4.3 Improving Vocabulary

Analysis reveals that 13% of all errors are due to the fact that a word is out of vocabulary. The out of vocabulary problem by itself can be solved by applying language models that use sub-word units instead of whole words. Although this approach solves out of vocabulary issue, it does not yield an improvement in terms of word error rate (Salimbajevs et al., 2015.). This time authors try the more obvious solution to deal with this problem and increase the training corpus used to prepare language model. In our case the training corpus was increased up to 47 million sentences with a 2.8 million unique word vocabulary, which reduced the out of vocabulary rate to 0.7%.

This resulted in WER reduction of 0.86%, and 61% of previously out of vocabulary words were correctly recognized. However, such a large increase in language model and vocabulary size resulted in the language model perplexity increasing on testing utterances by 647 compared to 498 obtained with language model trained on 22M sentence corpus and using 200K unit vocabulary, so the WER reduction was not as great as anticipated.

4.4 Evaluation

Combining all of the above mentioned improvements resulted in an improved final ASR system, which was then evaluated in terms of WER on both development and test sets (see Table 5).

System	Dev Set	Test Set
Baseline	18.06%	30.94%
Final	15.90%	28.43%

Table 5: WER of the final system.

5 Discussion

The Latvian language is an inflective language with complex morphology. Latvian also has a selection of prefixes and suffixes that can modify nouns, adjectives, adverbs and verbs. Because of this, two big problems arise: (1) high OOV rate, (2) errors in word endings.

Our error analysis reveals that endings contribute to 41% of errors and OOV words directly or indirectly cause 18% of errors. Together these two problems cause 59% of errors. Solving these problems will be very important for the further development of Latvian ASR.

However, not all errors are equal. Our results show that only 47% of errors make utterances difficult or impossible to understand. In most cases it is easy for a human reader to recover from errors in word endings, while in cases of OOV 82% of errors significantly change the meaning of the utterance.

We also found out that adjectives and verbs are more difficult to recognize than other parts of speech (excluding prepositions and particles). This is due to fact that they have the most variants of endings.

Non-canonical pronunciation can cause significant problems for ASR. In our development set no fractional numbers were recognized correctly because the pronunciation of the word “komats” (comma) was not canonical. We managed to reduce these errors by 50% by adding an alternative pronunciation to the G2P vocabulary.

Evaluation results show that there is a big difference in WER between the development and test sets. It seems that our random splitting was not very successful and resulted in uneven distribution of utterances which are hard to recognize. Both sets should have been tested before any analysis began. It is possible that some large class of errors was not identified.

Nevertheless, the final system showed noticeable improvement and outperformed the baseline system by about 2% WER on both test sets. This shows that our improvements were effective. Also it can be concluded that “surface” analysis of errors can help to improve speech recognition

6 Conclusions

In this paper we presented a surface error analysis of the Latvian Large Vocabulary Automatic Speech Recognition system.

The results show that more than 50% of errors are OOV and misrecognized word endings. Both of these problems are caused by the inflective nature and complex morphology of Latvian. Finding solution to these problems will greatly reduce the WER of Latvian ASR.

This analysis was then used to improve the present ASR system. After the changes the system showed 2% WER improvement on both the development and test sets.

In future we plan to perform more in-depth error analysis of errors in word endings. It is also important to find effective way of dealing with OOV, instead of just continuing to increase the size of the vocabulary.

7 Acknowledgments

The research leading to these results has received funding from the research project “Information and Communication Technology Competence Centre” of EU Structural funds, contract nr. L-KC-11-0003 signed between the ICT Competence Centre (www.itkc.lv) and the Investment and Development Agency of Latvia, research No. 2.4 “Speech recognition technologies”.

References

Darģis, R., & Znotiņš, A. (2014). Baseline for Keyword Spotting in Latvian Broadcast Speech. In *Human Language Technologies – The Baltic Perspective* (pp. 75–82). IOS Press.

Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, *29*, 137–158. doi:10.1016/S0167-6393(99)00035-7

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, *52*, 181–200. doi:10.1016/j.specom.2009.10.001

Goryainova, M., Grouin, C., Rosset, S., & Vasilescu, I. (2014). Morpho-Syntactic Study of Errors from Speech Recognition System. In N. C. (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Miao, Y., Zhang, H., & Metze, F. (2014). Towards speaker adaptive training of deep neural network acoustic models. *Proc. Interspeech*.

Oparin, I., Lamel, L., & Gauvain, J.-L. (2013). Rapid development of a Latvian speech-to-text system. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 7309–7313). doi:10.1109/ICASSP.2013.6639082

Pappu, A., Misu, T., & Gupta, R. (2014). Investigating Critical Speech Recognition Errors in Spoken Short Messages. In *Proceedings of the 5th International Workshop on Spoken Dialog Systems (IWSDS)*. Napa, California.

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)* (pp. 1547–1553).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Salimbajevs, A., & Pinnis, M. (2014). Towards Large Vocabulary Automatic Speech Recognition for Latvian. In *Human Language Technologies – The Baltic Perspective* (pp. 236–243). IOS Press.

Schwartz, R., Nguyen, L., Kubala, F., Chou, G., Zavaliagkos, G., & Makhoul, J. (1994). On Using Written Language Training Data for Spoken Language Modeling. In *Proceedings of the Workshop on Human Language Technology* (pp. 94–98). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1075812.1075830

Shinozaki, T., & Furui, S. (2001). Error analysis using decision trees in spontaneous presentation speech recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.* doi:10.1109/ASRU.2001.1034621

Stoyanchev, S., Salletmayr, P., Yang, J., & Hirschberg, J. (2012). Localized detection of speech recognition errors. In *Spoken Language Technology Workshop (SLT), 2012 IEEE* (pp. 25–30). doi:10.1109/SLT.2012.6424164

Vasilescu, I., Adda-Decker, M., & Lamel, L. (2012). Cross-lingual studies of ASR errors: paradigms for perceptual evaluations. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Do?an, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Salimbajevs, A. and Strigins, J. (2015) Using sub-word n-gram models for dealing with OOV in large vocabulary speech recognition for Latvian. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*