The 2011 KIT QUAERO Speech-to-Text System for Russian

Yury Titov[1], *Kevin Kilgour*[1,2], *Sebastian Stüker*[1,2], *and Alex Waibel*[1]

[1]Institute of1 Anthropomatics
[2]Research Group 3-01 'Multilingual Speech Recognition'
Karlsruhe Institute of Technology
Karlsruhe, Germany

{ytitov|kevin.kilgour|sebastian.stueker|alexander.waibel}@kit.edu

Abstract

This paper describes our current speech-to-text system for Russian that we are developing within the Quaero program. The system uses two different front-ends for obtaining different acoustic models that are used for system combination and cross-adaptation within a multi-stage decoding set-up. The acoustic model have been trained on manually transcribed data as well as untranscribed data, the language model on large amounts of data obtained from different sources. An error analysis shows the influence of pecularities of the Russian language on the word error rate, such as its morphological structure and specific orthorgaphic properties. Our system achieves a word error rate of 19.3% on the official Quaero 2010 evaluation set.

1. Introduction

This paper describes our current speech-to-text system for Russian that we are developing within the Quaero program.

Russian, like many other Slavonic languages offers various challenges to automatic speech recognition. While on the acoustic modelling side, the current standard models can be easily applied, Russian differs from many other languages, e.g. English, in two important ways when considering statistical language modelling.

Firstly, Russian is a highly inflected language—almost all content words have several inflections (word-endings) which change the grammatical case, gender, number, etc. of the word. As a consequence of this, the vocabulary of a Russian system needs to be be an order of magnitude larger than, e.g., for English in order to achieve similar out-of-vocabulary (OOV) rates.

Secondly, Russian shows a word order that is significantly relaxed compared to other languages [1]. While in practice a completely free ordering of words is not observed, and regular stylistic patterns are seen [1], [2], changes in word order frequently occur, generally to lend more weight to particular words in the sentence [1]. This fact generally allows building LVCSR systems for Russian using state-of-the-art n-gram language models, however showing higher perplexities on average.

Our system uses a multi-pass decoding strategy that combines systems with different front-ends and performs unsupervised speaker adaptation between passes. The models were trained on manually transcribed data provided by the Quaero program and untranscribed data collected from the World Wide Web (WWW).

In addition to the system itself we also provide an analysis of the errors committed by the system, showing the influence of the inflective nature of Russian on the word erro rate, as well as other language specific factors, such as peculiarities in the orthographie.

1.1. The Quaero Speech-To-Text Task

Quaero (http://www.quaero.org) is a French research and development program with German participation. It targets to develop multimedia and multilingual indexing and management tools for professional and general public applications such as the automatic analysis, classification, extraction, and exploitation of information. Also included in Quaero is basic research in the technologies underlying these application areas, including automatic speech recognition (ASR). Russian is included in the list of languages addressed by Quaero. We conducted our experiments in this paper on the official development (dev2010) and test (test2010) sets provided within the Quaero Program for the 2010 evaluations.

1.2. Related Work

ASR for Russian has received comparatively little attention in the literature (the first reported large-vocabulary recogniser for Russian appeared only in [3]). However, much work has been conducted in recent years on the language modeling ([4]) and acoustic modelling techniques ([5], [6]) for the speech recognition of Russian ([7], [8], [9]). This process was strongly supported in related fields. e.g. by the creation of various Russian audio corpora for training ([10], [11], [12], [13]) and progress in research of the various computational linguistics problems specific for Russian ([14], [15], [16]). Our recognition system has been developed out of our GlobalPhone speech recognition system for Russian [17] and is structured similar to our other ASR systems developed for TC-STAR and Quaero [18, 19].

2. Acoustic Model

We trained two sets of acoustic models, using two different acoustic front-ends. Both models consist of semi-continuous generalized quinphone models that use 16,000 distributions over 4,000 codebooks. Clustering of the quinphones was done with the help of a deci-

sion tree that asks questions about the phonetic properties of the phonemes in the context of the quinphones. Our HMM based acoustic model uses a phoneme set of 51 phonemes which closely correspond to the alphabet of the Russian language plus soft consonants. All phonemes are modelled with left-right Hidden Markov Models (HMMs) without state skipping and three states.

2.1. Acoustic Pre-Processing

For our system we used two acoustic front-ends: one based on the traditional Mel-frequency Cepstral Coefficients (MFCC) and the other basd on the warped minimum variance distortionless response (WMVDR) [20].

For the MFCC front-end, we extracted power spectral features using an FFT with a 10 ms frame-shift and a 16 ms Hamming window from the 16 kHz audio signal. We then computed 13 Mel-Frequency Cepstral Coefficients (MFCC) per frame. The MVDR front-end replaces the Fourier transformation by a warped MVDR spectral envelope which is a time domain technique to estimate an all-pole model using a warped short time frequency axis such as the Mel scale. For the MVDR front-end we used a model order of 30 without any filter-bank since the warped MVDR already provides the properties of the Mel-filterbank, namely warping to the Mel-frequency and smoothing.

Both frond ends apply vocal tract length normalization (VTLN) [21]. For MFCC this is done in the linear frequency domain, for MVDR in the warped frequency domain. Also, for both front-ends we performed cepstral mean subtraction and variance normalization on a per-utterance base. In order to incorporate dynamic features, for both front-ends fifteen adjacent frames were combined into one single feature vector. The resulting feature vectors were then reduced to 32 dimensions for the MFCC front-end and 42 for the MVDR front-end by using *linear discriminant analysis* (LDA).

2.2. Training Data

The training data for the acoustic model consists of two parts. The first part is the official QUAERO training dataset 2010 and 2011 which consists of approxomately 80 hours of manually transcribed audio data, including segmentation and annotation of speaker identities. The second part consists of broadcast news videos published by Channel One Russia on their website http://www.1tv.ru. Transcriptions for the videos were derived from the approximate transcripts available at the Channel One website. To do so, we first automatically transcribed the videos using our 2009 Russian evaluation system from Quaero with a language model that was biased towards the approximate transcripts. After that we extracted the word sequences of three words and more from the automatic transcriptions that have an exact match in the approximate transcripts. In that way we obtained 200h of transcriptions for the 1500h of videos that we had downloaded.

2.3. Training Procedure

The acoustic model training consisted of incremental splitting of Gaussians training, followed by 2 iterations of Viterbi training. The splitting of Gaussian training created codebooks with up to 128 Gaussian components per model. For all models we used one global semi-tied covariance (STC) matrix after LDA [22] as well as Vocal Tract Length Normalization (VTLN). In addition to that feature space constraint MLLR (cMLLR) [23] speaker adaptive training (SAT) was applied on top. We improved the acoustic models further with the help of Maximum Mutual Information Estimation (MMIE) training [24]. We applied MMIE training to the models after the 2 viterbi iterations, and to the models after the cMLLR SAT training, taking the adaptation matrices from the last iteration of the maximum likelihood SAT training and keeping them unchanged during the MMIE training.

3. Language Model

The language model of a Russian ASR system has to cope not only with a multitude of inflections but also with a loose word order, problems which are just not encountered when building an English or Spanish language model. We addressed this problem by vastly increasing the size of our language model from 14 million 4grams in the original system to 73 million 4grams in the final so called *very large language model*.

3.1. Text Data Sources

The Russian very large language model is build from 4 types of text data, broadcast news (BN), web data, books, and audio transcripts. The RU BN text data was collected by grabbing the top 50 websites in the mass-media category of the Yandex search engine's websites catalog (http://catalog.yandex.ru/yca/cat/Media/). Web resources used in dev/eval sets were excluded from the list for the text data collection. The Moshkov's open library (http://www.lib.ru/) provides a large collection of books from both Russian authors and foreign authors (translated into russian). The Quaero training texts consist of various web and BN texts scraped from the Internet. Also provided by Quaero are transcripts form 50 hours of audio data (Quaero 2010 dataset) which we, at the level of sentences, randomly divided into a tuning set and a training set. A detailed description of the corpora used to build the very large language model for Russian is described in Table 1. The total amount of the running words in the final text corpus is about 1.9 Billion words.

3.2. Text Data Normalization

After collecting the text data, most from online sources, text normalization was performed. It included so called yofication and replacement of the typos according the highest frequency candidates.

Yofication is a spellchecking procedure with an automatic spellcheker, in our case it was the Hunspell, v.1.2.7 with Lebedev's dictionary, for words in which the letter е, pronounced *"ye"*, should be replaced with the letter ё, pronounced *"yo"*. Since dictionary based spellchekers such as the Huspell cannot be used for the correction of the typos in named entities and acronyms, our text normalization

source	word count
Quaero Training Texts 2010	$\approx$ 110 mln
Books (Russian authors), www.lib.ru	$\approx$ 691 mln
Books (foreign authors), www.lib.ru	$\approx$ 890 mln
RU BN text data	$\approx$ 300 mln
Quaero 2010 audio data transcripts	$\approx 6k$

Table 1: *Description of the corpora used for the very large language model for Russian.*

assums that the words with typos take place more rarely than their correct spelling in the text corpra. Using this method approximately 259 thousand spelling mistakes were found in the corpus. As this normalization method could potentially corrupt some words which exist in both spelling variants a future implementation may include a manually created whitelist of words not to *correct*.

3.3. Language Model Training and Evaluation

Using our 500k vocabulary 4gram case sensitive language models were built for each of our Russian text sources. All bi- and tri-grams as well as all 4-grams occuring more than twice were included in the language model with modified Knesser-Ney smoothing. This was done using the SRI Language Modelling Toolkit [25]. The interpolation weights of the individual language models were calculated using the tuning text extraced from the transcripts of the Quaero 2010 audio data. The final LM contains 73 million 4grams and a vocabulary of 500k words with an out-of-vocabualary rate of 1.48% on the dev2010 set and 0.78% on the test2010 set. Detailed perplexity scores over the various Quaero datasets can be seen in the Table 2.

Language Model Name	Test09	Dev10	Test10
LM RU 2011	**253**	300	**232**

Table 2: *Perplexity scores of the final language model*

4. Decoding Strategy and Results

The final systems first segments the incoming audio into sentence like and chunks and performs speaker clustering on the resulting segments. Then a two stage decoding strategy with our four acoustic models is applied, as shown in Figure 1.

For the purpose of segmentation we performed a fast decoding pass on the unsegmented input data in order to determine speech and non-speech regions. Segmentation was then done by consecutively splitting segments at the longest non-speech region that was at least 0.3 seconds long. The resulting segments had to contain at least eight speech words and had to have a minimum duration of six seconds.

In order to group the resulting segments into several clusters, with each cluster, in the ideal case, corresponding to one individual speaker we used a hierarchical, agglomerative clustering technique which is based on the TGMM-GLR distance measure and the Bayesian Information Criterion (BIC) stopping criteria [26]. The resulting speaker labels were used to perform acoustic model adaptation in our the multi-pass decoding strategy.

In the first stage of the decoding that follows then, we first apply the VTLN trained acoustic models with the MFCC, MVDR front-end respectively. During decoding we use incremental VTLN and cMLLR for speaker adaptation. The result from this first stage is then combined using confusion network combination (CNC) [27].

For the second stage the cMLLR-SAT models are adapted on the CNC output from the first stage, using VTLN, cMLLR, and MLLR. Then decoding is performed with the adapted models and their results are then again combined by CNC. On top of this our yofication procedure was applied.

The same language model and vocabulary were used for all decodings. For the decodings in the second stage the frame shift was reduced to 8ms.

5. Error Analysis

The performance of the system evaluated using WER. The results, which can be seen in table X show that about two thirds of the errors are substituons. We performaned a error analysis identified to major types of substituon errors, Yo-homonyms and inflections.

5.1. Yo-homonym Errors

As discussed in section 3.2, yofication with a spellchecker was used to normalize the text data for the language model training. The same process was applied after the final decoding step. Even so, an analysis of the most frequent confusion pairs (table 3) reveal that the yofication procedure is far from perfect, with word pairs such as всё/все, чём/чем, Тягачёва/Тягачева being amoung the most common substitution errors. This is caused by the presence of so-called yo-homonyms, [28], which are the words that exist in both variants in the Russian. Addressing this problem will require some more sophisticated text normalization approaches as the usage of syntactical [28] or contextual information features [29].

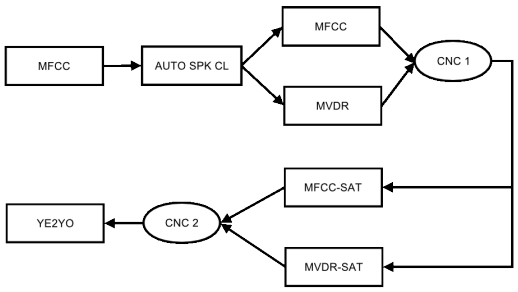

Figure 1: Decoding strategy of the Russian evaluation system

Number	Dev2010		Test2010	
	Frequency, tokens	Substitution Pair (REF/HYP)	Frequency, tokens	Substitution Pair (REF/HYP)
1	10	всё/все	18	всё/все
2	8	ну/но	16	ну/но
3	8	есть/и	6	эта/это
4	7	Нагайдели/Ногаидели	6	Тягачёва/Тягачева
5	6	Киворковой/Кеворковой	5	они/не
6	6	чтобы/что	5	её/его
7	6	эта/это	4	чтобы/что
8	5	таки/все-таки	4	который/которые
9	5	её/его	4	и/или
10	5	мне/не	4	считают/считает
11	5	которая/который	4	мне/не
12	5	Коротницкий/Каратницкий	4	это/эта
13	4	в/на	4	Си/Би-би-си
14	4	и/или	4	был/было
15	4	Гейц/Гейц	4	шестидесятипятилетию/шестидесяти
16	4	из/и	4	которая/которые
17	4	я/и	4	дело/дела
18	4	её/и	3	амнистия/амнистии
19	4	свободаньюс/ньюс	3	Хиллари/Хилари
20	4	них/не	3	Си/БиБиСи

Table 3: *The top 20 frequent substitutions ranking after the final decoding step for the Quaero Development Set 2010 (Dev10) the Quaero Evaluation Set 2010 (Test10)* . The word pairs are presented in the REF/HYP format: first word is a word from the manually transcribed reference(REF), a second is the recognized (hypothesized) word by the ASR system (HYP).

5.2. Inflection Errors

A second language-specific error source, that should be highlighted, are substitution errors caused by recognizing the correct word with the inflection; эта/это, амнистия/амнистии and который/которые (see table 3) are examples of this type of error. We should pay specific attention to these errors because they could indicate deficiencies in both the acoustic model, many inflections sound similar (while still being accusticaly distinguishable) as well as the language model, some inflections are accusticaly identicall and context knowledge is required. An automatic inflection detector was built with the aid the Mikhail Korobov's pyMorphy library to estimated how many of our errors are of this typ. For every substitution pair all possible inflections of a reference word were generated, if the hypothosis word was on the list of generated inflections, then that particular confusion pair was considered an inflection error. 20% of the subistution errors in the dev2010 set were inflection errors and in the test2010 set inflection errors were responsible for 15% of the subistution errors.

6. Discussion and Future Work

To develop an effective speech recognition system for a Slavonic languages (and in particular for Russian) it is necessary to solve some difficulties concerning the peculiarities of these languages. These languages belong to the category of synthetic languages, which are characterized by a tendency to combine (synthesize) lexical morphemes (or several lexical morphemes) and one, or several, grammatical morphemes into one word-form [9]. The rich morphology results in a large search vocabualy containing many accoustically similar words. Endings especially are also often droped in fast or conversational speech which makes the ASR system's task of recognizing

Step	Test10, ci	Test10, cd	Dev10, ci	Dev10, cd
MFCC	22.88	23.89	24.98	25.93
MVDR	24.64	25.68	27.90	28.85
CNC1	21.70	22.64	24.41	25.29
MFCC-SAT	20.72	21.80	23.02	23.97
MVDR-SAT	19.81	20.91	22.49	23.47
CNC2	19.30	20.29	21.93	22.83
YE2YO	**19.27**	20.26	**21.90**	22.80

Table 4: Detailed final system performance over the QUAERO development and test 2010 datasets with case-insensitive (ci) and case-dependant (cd) scores, Word Error Rate (WER), %

the correct words ever harder. To reduce the confusions caused by this the authors would like to incorperate the modern approaches in unsupervised morphological segmentation [30]. The second specifics of Russian is, the loose word order in Russian sentences. It complicates the creation of statistical language models based on n-grams as well as grammars, and decreases their effectiveness. Recent research suggest using neural networks fpr language modelling to address this problem ([31],[32]). Future systems should also consider evaluating alternative acoustic front-ends like PLP and M-RASTA filters.

7. Conclusions

In this paper we present our Russian LVCSR system and its application to the Quaro 2010 evaluation task containing an even mixture of conversational speech and broadcast news. We demontrate that with a state-of-the art acoustic model trained on transcribed and untranscribed data we are able to achieve WER rates of around 20% on a 2nd pass after speaker adaptation. A confusion network combination of systems with different acoustic front-ends (MFCC and MVDR) reduces the WER further. To combat the loose word order in Russian we employ a large n-gram language model built from about 1.9 Billion words. Our error analysis shows that a large amount of the remaining errors are caused by inflections and Yo-homonyms.

8. Acknowledgement

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. 'Research Group 3-01' received financial support by the 'Concept for the Future' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

9. References

[1] N. D. Andreeva, Ed., *Statistical and Combinatorical Modeling of a Natural Language*, Nauka, 1965.

[2] E.W.D. Whittaker and P.C. Woodland, "Comparision of language modelling techniques for russian and english," in *ICSLP*, 1998.

[3] D. Kanevsky, M. Monkowski, and J. Sedivy, "Large vocabulary speaker-independent continuous speech recognition in russian language," in *SPECOM*, 1996.

[4] M. Zulkarneev, "An approach to compensation for language modeling errors in the key-spotting systems," in *SPECOM*, 2004.

[5] I. Kipyatkova and A. Karpov, "Creation of multiple word transcriptions for conversational russian speech recognition," in *SPECOM*, 2009.

[6] Andrey L. Ronzhin, Rafael M. Yusupov, Izolda V. Li, and Anastasia B. Leontieva, "Survey of russian speech recognition systems," in *SPECOM 2006*, June 25-29 2006.

[7] J. Psutka, J. Hajic, and W. Byrne, "The development of asr for slavic languages in the malach project," in *ICASSP*, 2004.

[8] J. Psutka, P. Ircing, J.V. Psutka, J. Hajic, W. J. Byrne, and J. Mirovsky, "Automatic transcription of czech, russian, and slovak spontaneous speech in the malach project," in *INTERSPEECH*, 2005.

[9] A.L. Ronzhin and A.A. Karpov, "Large vocabulary automatic speech recognition for russian language," in *Second Baltic Conference on Human Language Technologies*, 2005.

[10] N.V. Bogdanova, A.S. Asinovsky, M.V. Rusakova, A.I. Ryko, C.B. Ctepanova, and T. Yu. Sherstinova, "Speech corpus as a tool for monitoring and fixation of various forms of natural language," in *Dialog21*, 2010.

[11] S.B. Stepanova, A.S. Asinovsky, A.I. Ryko, and T.Yu. Sherstinova, "Phonetic realization of russian inflections in the ord speech corpus of everyday communication," in *Dialog21*, 2010.

[12] A. Asinovsky, N. Bogdanova, M. Rusakova, A. Ryko, S. Stepanova, and T. Sherstinova, "Speech corpus of russian everyday communication "one speaker's day" (the ord corpus)," in *SPECOM*, 2009.

[13] E. Lyakso and O. Frolova, "Russian infants and children's sounds and speech corpuses for language acquisition studies," in *INTERSPEECH*, 2010.

[14] O.Lyashevskaya and J. Kuznetsova, "Russian framenet: Towards a corpus-based dictionary of constructions," in *Dialog21*, 2009.

[15] A. Sokirko, "Bystroslovar: morphological prediction new russian words using very large corpora," in *Dialog21*, 2010.

[16] O. Lyashevskaya, I. Astafyeva, A. Bonch-Osmolovskaya, A. Garejshina, J. Grishina, V. Dyjachkov, M. Ionov, A. Koroleva, M. Kudrinsky, A. Lityagina, E. Luchina, E. Sidorova, S. Toldova, S. Savchuk, and S. Kovaly, "Nlp evaluation: Russian morphological parsers," in *Dialog21*, 2010.

[17] Sebastian Stüker and Tanja Schultz, "A grapheme based speech recognition system for russian," in *Proceedings of the 9th International Conference "Speech And Computer" SPECOM'2004*, Saint-Petersburg, Russia, September 2004, pp. 297–303, Anatolya.

[18] Sebastian Stüker, Christian Fügen, Susanne Burger, and Matthias Wölfel, "Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006, ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 521–524, ISCA.

[19] Sebastian Stüker, Christian Fügen, Florian Kraft, and Matthias Wölfel, "The isl 2007 english speech transcription system for european parliament speeches," in *Proceedings of the 10th European Conference on Speech Communication and Technology (INTERSPEECH 2007)*, Antwerp, Belgium, August 2007, pp. 2609–2612.

[20] M.C. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectralestimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, September 2005.

[21] Puming Zhan and Martin Westphal, "Speaker normalization based on frequency warping," in *ICASSP*, Munich, Germany, April 1997.

[22] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," Tech. Rep., Cambridge University, Engineering Department, February 1998.

[23] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," Tech. Rep., Cambridge University, Engineering Department, May 1997.

[24] D. Povey and P.C. Woodland, "Improved discriminative training techniques for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, May 2001.

[25] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *ICSLP*, 2002.

[26] Q. Jin and T. Schultz, "Speaker segmentation and clustering in meetings," in *ICSLP*, 2004.

[27] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, October 2000.

[28] B. Lobanov, "The problem of the «yo»-homographs resolution in text-to-speech synthesis"," in *Dialog21*, 2009.

[29] Y. Zelenkov, I. Segalovich, and V. Titov, "Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов," in *Dialog21*, 2005.

[30] M. Creutz and K. Lagus, "Inducing the morphological lexicon of a natural language from unannotated text," in *AKRR'05*, 2005.

[31] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, 2010.

[32] T. Mikolov, J. Kopecky, L. Burget, O. Glembek, and J. Cernocky, "Neural network based language models for highly inflective languages," in *ICASSP*, 2009.