Transcription of Russian Conversational Speech

Lori Lamel[1,2], *Sandrine Courcinous*[2], *Jean-Luc Gauvain*[1,2], *Yvan Josse*[2] *and Viet Bac Le*[2]

[1]Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
{lamel,gauvain}@limsi.fr

[2]Vocapia Research
3, rue Jean Rostand
91400 Orsay, France
{courcinous,josse,levb}@vocapia.com

Abstract

This paper presents initial work in transcribing conversational telephone speech in Russian. Acoustic seed models were derived from other languages. The initial studies are carried out with 9 hours of transcribed data, and explore the choice of the phone set and use of other data types to improve transcription performance. Discriminant features produced by a Multi Layer Perceptron trained on a few hours of Russian conversational data are contrasted with those derived from well-trained networks for English telephone speech and from Russian broadcast data. Acoustic models trained on broadcast data filtered to match the telephone band achieve results comparable to those obtained with models trained on the small conversation telephone speech corpus.

1 Introduction

This paper reports on recent research aimed at developing a speech-to-text transcription (STT) system for conversational telephone speech (CTS) for the Russian language. A survey of Russian speech recognition systems can be found in [1]. Traditionally STT systems are trained on large amounts of carefully transcribed speech data and huge quantities of written texts. However obtaining the needed transcribed audio data remains quite costly and requires substantial supervision. This is particularly onerous for conversational speech where the collect and transcription are much more complicated and time consuming than for broadcast data.

Several research directions have addressed reducing these costs [2] and there has been growing interest in producing and training with audio data that are associated with quick transcriptions [3]. It is possible to find associated texts for some audio sources, this type of data is not very close to conversational speech. A variety of lightly and unsupervised acoustic model approaches have been proposed, most relying on supervision from a language model. The approaches differ in their details: use or not of confidence factors [4] or [5], [6], doubling vs iterative training [7] and the amount of data used. Collecting text from the Web was used by [8] to improve STT performance on conversational telephone speech in Mandarin and English.

In the context of the Quaero program (www.quaero.org), LIMSI and Vocapia Research developed a speech-to-text transcription systems for broadcast data in Russian [9]. Since only a 9-hour corpus of transcribed Russian conversational data was available for this work, some initial experiments were carried out using broadcast audio for acoustic model training.

The next section gives an overview of the characteristics of the Russian language, followed by a description the speech transcription system, and the corpora used in this study. This is followed by a description of the language models, the phone set and acoustic models, after which experimental results are provided.

2 Russian Language

Russian language belongs to Indoeuropean family, Slavic group, East Slavic branch and is written with a modern variant of the Cyrillic alphabet. Nouns, proper names, adjectives, pronouns, numerals and participles are subject to declension in six cases, two numbers and three genders. All of these can be stacked one upon the other, to produce multiple derivatives of a given word. Agglutinative compounds are also very frequent, over the last century, abbreviated forms are being used for some compounds.

According to www.internetworldstats.com, as of May 2011 Russian is the 9th most frequently used language on the Internet. With almost 60 million users, Russian users represent only 3% of the worldwide Internet users, but has had one of the fastest growth rates over the last decade.

The Russian pronunciation is almost phonetic, meaning that there is relatively straightforward correspondence between letters and sounds. Consonants are divided into palatalized (soft, also called 'wet') and non-palatalized (hard) ones. Stress is free and moving, it can fall on any syllable of the word and on different syllables of the word. The lack of stress information poses a challenge for pronunciation generation, since the stress position can modify the vowel pronunciation. Word order in Russian is free, and by changing the word order any word in a sentence can be emphasized.

3 Speech Recognizer Overview

The LIMSI conversational speech transcription system has two main components, an audio partitioner and a word recognizer. As described in [10], the conversational speech system was originally derived from a broadcast news system [11]. The word recognizer uses continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling. The audio partitioner uses an audio stream mixture model [12] to divide the acoustic signal into homogeneous segments, and associate appropriate speaker labels with the segments. Non-speech segments are detected and rejected using Gaussian mixture models representing speech and silence, after which an iterative maximum likelihood segmentation/clustering procedure is applied to the speech segments. The result of the procedure is a sequence of non-overlapping segments with their associated segment cluster labels, where each segment cluster represents one speaker.

Word recognition is performed in a single pass, which generates a word lattice with cross-word, position-dependent, gender-dependent acoustic models. This is followed by a lattice decoding using a 4-gram language model. Unsupervised acoustic model adaptation is performed for each segment cluster prior to decoding. The acoustic model adaptation relies on a tree organization of the phone model states to automatically create adaptation classes as a function of the available data.

Table 1: Russian vowels and other symbols. The highlighted boxes show the 4 generic vowels added to form the 50 and 55 phone sets.

vowels		
IPA	phoneme	example
a	a	альба
ɛ	e	августе
i	i	авгиевых
o	o	аудио
u	u	августе
^{j}a	à	австрия
^{j}ɛ	è	авгиевых
ɨ	ì	австрии
^{j}o	ò	ёжик
^{j}u	ù	австрию
ə	ø	августовские
ɐ	Â	поблагодарив
j	y	августовский

generic vowels		other symbols	
phoneme	represents	phoneme	example
O	o a ø	.	silence
E	e i ø	&	filler
À	à i yi ø	®	breath
È	è i ø ò		

4 Corpus

About 9 hours of Russian conversational telephone speech (CTS) data were available for this study. These were split into training and development sets. The training set is comprised of 44 conversations, with durations ranging from 2 minutes to 30 minutes, for an average length of about 11 minutes. 81% of the data is from female speakers, with 19% from male speakers. There are a total of about 69k words in the training set. About 1 hour of data was reserved for development purposes. The development data is comprised of 7 conversation sides, with 89%/11% female/male repartition and a total of 9.5k words. The training data consist of 2-channel telephone recordings, and the audio quality is quite varied, with some distortion for some recordings, but not much crosstalk.

5 Phone Sets & Pronunciations

Tables 1 and 2 show respectively the set of vowels and consonants used in the pronunciation lexicon and the acoustic models. The first phone set contained 51 elements, as used in the BN Russian systems. This included 35 consonants, 13 vowels and 3 non-speech units for breath, fillers and silence. As shown in Table 3, four contrastive phone sets were investigated. In addition to the 51 phone set, one removes the 5 soft consonants which are relatively rare in the CTS data and a second introduces 4 generic vowels to address the challenge of pronunciation generation with uncertain stress. In this case a single generic vowel is used to represent several possible phone realisations. The final 55 phone set containing both the soft consonants and the generic vowels was not used in the initial experiments.

The pronunciation lexicons for each of the phone sets were created principally using grapheme-to-phoneme conversion rules since as mentioned earlier this is relatively straightforward for the Russian language. Some pronunciations were obtained from the Master dictionary developed for the Quaero broadcast data STT system. For the 46 and 51 phone sets, a morphological analyzer [13] for the Russian language was modified to return the stress for a word in context. Using this information, the appropriate lexical-stress dependent phonetization can be generated taking into account vowel reduction. Using the rules summarized in Table 5 (after [14]), it is possible to model the vowel reduction affecting the vowels /aɛiou/. Depending upon the stress, these can be mapped to one of their counterparts vowels /əeiɐ/. These rules can generate multiple pronunciations for a word, resulting in an average of 1.6 pronunciations/word for the 46 and 51 phone lexicons. Since the 50 and 55 phone sets use the generic vowel, they have essentially only one pronunciation/word.

Table 2: Russian consonants. The highlighted boxes show the 5 infrequent soft consonants removed to form the 46 and 51 phone sets.

hard consonants			soft consonants		
IPA	phoneme	example	IPA	phoneme	example
b	b	альба	b^j	B	айсберг
d	d	адовы	d^j	D	андерс
f	f	августовские	f^j	F	обувь
g	g	авгиевых	g^j	G	ангел
k	k	августовские	k^j	K	арктике
ɫ	l	али	λ	L	алекс
m	m	августовским	m^j	M	атоме
n	n	адресной	n^j	N	айне
p	p	альпы	p^j	P	опера
r	r	австрии	r^j	R	авторе
s	s	августе	s^j	S	айресе
t	t	августовские	t^j	T	августе
v	v	авгиевых	v^j	V	авторстве
x	x	авгиевых	x^j	X	анхель
z	z	азией	z^j	Z	образе
ʦ	§	акцией	$ʧ^j$	¢	алчность
ʃ	ç	аншлюса	$ʃ^j$	Ç	атомщики
			ʒ	J	ужас

6 Acoustic Modeling

Two sets of acoustic features were used in this work. The first are PLP-like [15] features and second are probabilistic features produced by Multi Layer Perceptron (MLP) [16]. Previous experiments with alternate MLP features have shown that the TRAP-DCT features [17] have comparable performance to the warped linear predictive temporal patterns (wLP) but are much cheaper to obtain.

For the PLP features, 39 cepstral parameters are derived from a Mel frequency spectrum, with cepstral mean removal and variance normalization carried out on a segment-cluster basis, resulting in a zero mean and unity variance for each cepstral coefficient [11]. The TRAP-DCT features are obtained from a 19-band Bark scale spectrogram, using a 30 ms window and a 10 ms offset. A discrete cosine transform (DCT) is applied to each band (the first 25 DCT coefficients are kept) resulting in 475 raw features, features which are the input to a 4-layer MLP with the bottleneck architecture [18]. The size of the third layer (the bottleneck) is equal to the desired number of features (39). A 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is combined with the other features, resulting in a total of 42 (plpf0) or 81 parameters (mlpplpf0).

Table 3: Different Russian phone sets investigated.

#phones	46	51	50	55
30 consonants	x	x	x	x
5 soft (wet) consonants		x		x
13 vowels	x	x	x	x
4 generic vowels			x	x
3 non-speech symbols	x	x	x	x

Table 4: MLP cross-validation frame accuracies.

Languages	#MLP targets	CV Accuracy
US CTS	108	48.09 %
Ru BN_{tel}	147	45.07%
Ru CTS	144	47.81%

Features derived from three different MLP networks were used. The first that was trained on over 2000 hours of conversational telephone speech for US English. The second was trained on Russian broadcast data that was filtered to match the telephone bandwidth (Ru BN_{tel}). The third was trained on the Russian CTS data. All networks were trained using the scheme proposed in [19], reserving a portion of the data for cross-validation to monitor performance (shown in Table 4. The MLP targets, correspond to the individual states for each phone and one state for each of the pseudo phones (silence, breath, filler).

As in [11] the acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (typically 32 components). The triphone-based phone models are word independent, but word position-dependent. The

Table 5: Vowel reduction rules after [14].

		initial	other pre-stress	first pre-stress	stress	post-stress
a, o	after hard consonant except Post alveolar	/a/	/ə/	/ɐ/	/a/ or /o/	/ə/
я, e	after sweet consonant	ø	/ə/	/i/ or /ɛ/	/a/ or /ɛ/	ø
a, e, o	after hard Post alveolar	ø	/i/ or /ə/	/ɛ/ or /i/	/a/ or /ɛ/ or /o/	/ə/ or /i/

states are tied by means of a decision tree to reduce model size and increase triphone coverage. State-tying constructs one tree for each state of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states. Silence is modeled by a single state with 1024 Gaussians.

The most frequent phone contexts in the training data are modeled, with separate cross-word and word-internal statistics. Gender-dependent acoustic models were built using MAP adaptation of the gender-independent models [20]. Depending upon the number of phones, the acoustic models cover between 6000 to 6380 phone contexts with about 2600 tied states. The sets of questions used by the divisive tree based clustering algorithm concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones. Depending upon the phone set, the number of questions ranges from 73 to 92.

7 Language Modeling

The methods used for text normalization and vocabulary section for broadcast data were applied. Basic normalization rules were applied to the Russian texts extracted from the HTML web pages. In Russian, like in other languages, abbreviations are common. These abbreviations were replaced by their expanded forms. Foreign words in Russian texts (which appear mainly newspaper articles) can be written with Latin or Cyrillic characters. It was decided to remove all the Latin words since none of them were observed in the CTS transcriptions used for this experiment. Cyrillic words that are completely written with uppercase letters were processed as acronyms.

Regular expressions were used to detect numbers and capture their right and left contexts. Depending on these contexts, the numbers were classified into one of the following categories: cardinal, ordinal, time, amount, phone number. The category was used when converted the number in a spoken form. Number declensions were not taken into account at this stage of work, so each number is written in the masculine nominative form.

After the punctuation was tokenized, the texts were formatted one sentence per line, and then the punctuation markers were removed. Only minor normalizations were applied to the manual transcriptions of broadcast and CTS data.

Table 6: Summary of the Russian text corpora used for language model training.

Source	Nature	Words	Vocabulary
Texts	Articles	296.1M	1.8M
BN	transcripts	741K	74K
CTS	transcripts	69K	8.9K
OpenSubtitles	subtitles	483K	53K
Total		297.3M	1.8M

Table 7: Characteristics of the component Russian language models estimated on the subsets of the available training texts.

Source	PPX	#4-grams	Interpol. coeff.
Texts	2371.1	188M	0.22
BN transcripts	3227.8	638K	0.02
CTS transcripts	1246.0	60K	0.69
OpenSubtitles	2715.5	325K	0.07

Table 6 summarizes the text corpora used for LM training. These are grouped in 4 subsets. Most of the texts are from written sources, and almost all of the audio transcripts are from broadcast data with only 69k words of transcripts of conversational telephone speech. The table specifies the total number of words for each subcorpus and the number of distinct words after normalization. There are almost 2M distinct word forms. The 1h dev set has 9K words, and was collected and transcribed in the same manner as the audio training corpus.

The recognition vocabulary was selected by interpolation of unigrams on the development data set to minimize the OOV rate on this data by selecting the most probable words. A set of 500K words was chosen that resulted in an OOV rate of just under 3% on the development data set.

Individual models are built using the Kneser-Ney smoothing algorithm and then interpolated. The interpolation coefficient are chosen automatically using the EM algorithm and targeting our development data set. Table 7 shows the characteristics of these individual models. After interpolation the resulting language model is pruned. The final interpo-

Table 8: Word error rate (%) using plpf0 AMs trained on only 8h of CTS data, with three different phone sets.

phone sets	%WER
46 GI	60.0
46 GD	58.9
51 GD	59.0
50 GD	58.2

lated LM contains 19M 4-grams and results in a dev data perplexity of 742.1.

8 Experimental results

To set a first baseline, the CTS development data were processed with the 2010 Russian Quaero broadcast news models [9]. A case-insensitive word error rate of 84.3% was obtained. The 51 phone set acoustic models from this system are discriminatively trained on about 100 hours of predominantly wideband broadcast audio data with transcripts, and use the 81 parameter mlpplpf0 feature vector. The language models are estimated on about 300 million words of varied texts and optimized for broadcast data. The word error rate of this system on a mix of broadcast news and broadcast conversation data used in the Quaero program is about 20%.

Table 8 summarizes a series of experiments carried out with plpf0 acoustic models with lexicons represented with three of the phone sets from Table 3, and the last LM in Table 9. As a reminder, the phone sets differ in as to whether they include the infrequent soft consonants and/or if they use generic vowels to cover the contextual, lexical stress dependent realization of vowels. The first two entries gives the WER using the smallest phone set with 43 phones and 3 non-speech units, using gender-independent (GI) and gender-dependent (GD) acoustic models. Since the GD models outperformed the GI ones all remaining results make use of GD models. The third entry makes use of the 51 phone set which includes the 5 soft consonants. There is essentially no change in performance, and since these phones are rare in the CTS training data, it was decided to not include them. The last entry uses the 4 generic vowel models instead of trying to generate word stress-dependent pronunciation variants via a morphological analysis. This latter phone set gives a slight performance gain relative to the other two sets. When pronunciation counts derived from their appearance in the training data are used with 46 and 51 phone set models, the WER is reduced by about 0.4% absolute.

Table 9 gives contrastive results in an effort to assess the contribution of the different text corpora on system performance. These experiments were carried out with plpf0 models trained on the 8 hour corpus, using the 50 phone set. A WER of 61% is obtained using a language model trained only on web texts, but targeting the dev data. Including either the BN transcripts or the OpenSubtitles results in the same WER reduction, with an additional improvement using both. Even though the amount of CTS transcripts is quite small, as can be expected these are an important LM component.

Table 9: Comparing LMs trained on different text corpora using Ru CTS 50 phone plpf0 models.

Train corpus	%WER
BN text	61.0
BN+BNtrans	59.9
BN+subtitles	59.9
BN+BNtrans+sub	59.4
BN+BNtrans+sub+CTStrans	58.2

Table 10: Comparing acoustic models trained on Russian CTS data with acoustic models trained on Quaero Russian broadcast data, using different sets of MLP features derived from Russian or US CTS data or Russian BN data filtered to telephone band.

AM trn	features	MLP trn	#phones	%WER
CTS data 8 hours	plpf0	-	50	58.2
	mlpplpf0	Ru CTS	50	50.7
	mlpplpf0	US CTS	50	51.2
	mlpplpf0	Ru BN$_{tel}$	50	51.0
	plpf0	-	51	59.0
	mlpplpf0	Ru CTS	51	52.0
	mlpplpf0	US CTS	51	52.7
	mlpplpf0	Ru BN$_{tel}$	51	51.7
Quaero BN data 195 hours	plpf0	-	51	60.0
	mlpplpf0	Ru BN$_{tel}$	51	51.6
	mlpplpf0	Ru CTS	51	51.5

Table 10 compares acoustic model training with three sets of mlpplpf0 features, The upper part gives results with AMs trained on CTS data, and the lower part AMs trained on BN data filtered to match the telephone band. Since the BN systems use a 51 phone set results are also given for the CTS trained models using the 51 phone set to have a more fair comparison. A first observation is that the mlpplpf0 models all significantly outperform plpf0 models, independent of which data were used to train the MLP network. While the models trained on the CTS audio data using the Ru CTS MLP features give the best performance, there is not a big difference if the other MLP feature sets are used, and it is slightly better to use the Ru BN$_{tel}$ MLP features than the US CTS ones. The WER with the filtered Russian BN training date is 51.5% which is not that much worse than the best result of 50.7% trained with only the Ru CTS data.

Unsupervised training: Given the high error rate of the best model, and the lack of additional transcribed resources first attempt to carry out acoustic model unsupervised training using the LDC RUSTEN corpus (LDC2006S34) was unsuccessful- using the system with a WER of about 50% to transcribe the 30 hours RUSTEN data and add it to the acoustic training data (without any confidence filtering as was done in [6] , or lattice-based training as proposed by [21]), degraded the performance by about 1% absolute.

After the experiments reported here were carried out, we obtained another 1 hour set of development data. This was recorded and transcribed completely independently of the first set. Although there was a loss on the first development set, there small improvement was obtained on the new dev data (from 62.0% to 59.0%), however the overall WER is higher on the new dev suggesting that there is a mismatch between the 3 data sources. Acoustic models with mlpplpf0 features, trained with the 50 and 55 phone sets, obtained comparable performance on both dev sets.

Morphological decomposition: A very simplistic morphological decomposition experiment was also carried out in an attempt to reduce the recognition vocabulary size and improve lexical coverage. This is much simpler than the method proposed in [22]. To do so a list of 50 declension suffixes was defined, and a single ending removed from the word as long as the remainder has at least 5 characters. Although the size of the recognition lexicon was reduced to 324k entries with a 20% reduction in OOVs, there was no reduction in the WER.

9 Conclusions

This paper has described our initial work in developing a speech-to-text transcription system for conversational Russian telephone speech data. Since only a small corpus of transcribed Russian CTS data was available, attempts were made to use other data sources. First attempts to use additional Russian CTS audio in an unsupervised manner and to use a simple morphological decomposition to reduce the OOV rate were not successful. Training acoustic models with broadcast data filtered to match the telephone band resulted in a relative word error rate quite comparable to that obtained with models trained on the CTS corpus. Discriminative features obtained with multi-layer perceptions, whether estimated on the CTS corpus or other available sources (US CTS or Ru BN_{tel}) when combined with plpf0 features gave significant improvements over the plpf0 features alone.

References

[1] A. Ronzhin, R. Yusupov, I. Li, A. Leontieva, "Survey of Russian Speech Recognition Systems," *SPECOM'06,* St. Petersburg, June 2006, 54-60.

[2] O. Kimball, C.L. Kao, R. Iyer, T. Arvizo, J. Makhoul, "Using quick transcriptions to improve conversational speech models," *ISCA Interspeech'04*, 2265-2268, 2004.

[3] C. Cieri, D. Miller, K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text," *LREC'04*, 69-71, 2004.

[4] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, H. Ney, "Cross domain automatic transcription on the TC-STAR EPPS corpus," *IEEE ICASSP'05*, **1**:825-828, 2005.

[5] M. Bisani, H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, **50**(5):434-451, 2008.

[6] F. Wessel, H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. on Speech & Audio Processing*, **13**(1):23-31, 2005.

[7] J. Ma, R. Schwartz, "Unsupervised versus supervised training of acoustic models," *Interspeech'08*, 2374-2377, 2008.

[8] I. Bulyko, M. Ostendorf, A. Stolcke, "Getting More Mileage fromWeb Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures," *HLT-NAACL 2003*, 3-7; 2003.

[9] L. Lamel et al., "Speech Recognition for Machine Translation in Quaero," *IWSLT'11*, San Francisco, Dec. 2011.

[10] S. Matsoukas et al., "Advances in Transcription of Broadcast News and Conversational Telephone Speech within the Combined EARS BBN/LIMSI System," *IEEE Trans. Audio, Speech & Language Processing*, **14**(5):1541-1556, 2006.

[11] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**:89-108, 2002.

[12] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**, 1335-1338, Sydney, Dec. 1998.

[13] "http://www.aot.ru/download/rus-src-morph.tar.gz."

[14] C. Meunier, *Grammaticalement correct russe ! Grammaire russe alphabétique*.

[15] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," *JASA*, **87**:1738-1752, April, 1990.

[16] P. Fousek, L. Lamel, J.L. Gauvain, "On the Use of MLP Features for Broadcast News Transcription," *TSD08*. LNCS 5246/2008, 303.10, Springer Verlag, Berlin, 2008.

[17] P. Schwarz, P. Matějka, J. Černocky, "Towards Lower Error Rates In Phoneme Recognition," *TSD'04*, 465-472, Brno, 2004.

[18] F. Grézl, P. Fousek, Optimizing Bottle-Neck Features for LVCSR, *IEEE ICASSP'08*, 4729-4732, Las Vegas, 2008.

[19] P. Fousek, L. Lamel, J.L. Gauvain, "Transcribing Broadcast Data Using MLP Features," *ISCA Interspeech'08*. 1433-1436, Brisbane, 2008.

[20] J.L. Gauvain, C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech & Audio Processing*, **2**:291-298, 1994.

[21] T. Fraga-Silva, L. Lamel, J.L. Gauvain, "Lattice-based Unsupervised Acoustic Model Training." *IEEE ICASSP'11*, Prague, 2011.

[22] A. Karpov, I. Kipyatkova, A. Ronnzhin,"Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis", *ISCA InterSpeech'11*, Florence, Aug 2011, 3161-3164.