Phone Durations Modeling for Livvi-Karelian ASR

Irina Kipyatkova(✉) and Ildar Kagirov

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 14th Line, 39, 199178 St. Petersburg, Russia
{kipyatkova,kagirov}@iias.spb.su

Abstract. This paper presents the results of experiments conducted during development of an automatic speech recognition system for the low-resource Karelian language (Livvi-Karelian dialect). The main issues addressed within this work are related to acoustic modeling, viz. the treatment of long and short phonemes. There are two approaches to modeling phonological duration in the so-called quantity languages: representation of long and short phonemes as distinct units, and interpretation of long phonemes as reduplicated. There is currently no consensus on which strategy is the most promising. The Livvi-Karelian case is further complicated by the fact that the phonology of Karelian was heavily influenced by Russian, so that a direct transfer of the methods applied to other Balto-Finnic languages is questionable. In the course of the study, experiments were conducted with both approaches, showing that treating long phonemes as reduplicated outperforms the approaches implying introduction of long and short counterparts in the phoneme set. The usage of alternative transcriptions for words with long consonants further improved the recognition accuracy. In addition, the present study contributes to the application of DNN approaches to the tasks of language and acoustic modeling in low-resource languages. In the future works, it is planned to improve the performance of the developed system with transfer techniques and advanced data augmentation procedures.

Keywords: Low-Resource Languages · Automatic Speech Recognition · Livvi-Karelian · Phoneme Duration Modeling

1 Introduction

Automatic speech recognition (ASR) systems play an important role in various domains, such as the development of voice assistants, speech-to-text applications and language learning tools. For a variety of languages, however, the accurate modeling of phoneme durations is crucial for ensuring high recognition accuracy, as the duration of phonemes can carry important linguistic information. The aim of this paper is to investigate and compare two distinct approaches for acoustic modeling in quantity languages (i.e., languages with phonemic distinction between long and short sounds): modeling long and short phonemes as separate units versus representing long phonemes as a sequence of two (or more) short phonemes. The research is conducted on the data from the low-resource Karelian language (Livvi-Karelian dialect).

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 87–99, 2023.
https://doi.org/10.1007/978-3-031-48312-7_7

Among the main tasks of this research are evaluating the accuracy of word recognition (WER metrics) when using separate models for long and short phonemes, and comparing this approach with modeling long phonemes as reduplicated units.

In the following sections of the paper, a detailed description of the current approaches to the problem is provided, the collected database and the experiments conducted are presented. The obtained results, including the analysis of the advantages and limitations of different approaches to modeling long and short phonemes in Livvi-Karelian ASR, is discussed among other things. In the conclusion, the research findings and their practical significance, as well as future work projects are outlined.

2 Related Work

2.1 Speech Recognition for Low-Resource Languages

Nowadays, there are two main approaches to development of ASR systems: "traditional" and end-to-end approaches. In the traditional approaches, ASR system is compound of several components: acoustic model (AM), language model (LM), and Pronunciation model (PM). The AM is responsible for mapping acoustic features of each frame to phonetic units, specifically phonemes. The LM associates the phoneme sequence generated by the AM with the sentence having the highest probability. On the contrary, in end-to-end ASR systems there is a single neural model transforming the speech signal to sequence of words [1–3]. Although end-to-end is a state-of-the-art approach showing better performance in terms of decoding speed, it typically requires large training data, and its performance has not surpassed that of traditional models in low-resource speech recognition tasks [4]. Thus, the end-to-end approach is not applicable to low-resource languages, that is, languages for which little data (regarding natural language processing tasks) exists by definition.

Currently, deep neural networks (DNNs) are extensively employed for training both acoustic and language models in ASR systems. For acoustic modeling DNNs are often combined with Hidden Markov Models (HMMs), thus forming hybrid DNN/HMM model. This approach has gained popularity due to its high performance in various applications. For instance, in [5], hybrid DNN/HMM acoustic models were employed for a Sinhala language ASR system. The results demonstrated that these models outperformed HMMs based on Gaussian Mixture Models (GMMs) by achieving a 7.48% improvement in word error rate (WER) on the test dataset.

In another study [6], experiments were conducted on multilingual speech recognition, focusing on low-resource languages including North American Cree and Inuit languages. The researchers investigated the use of factorized time delay neural networks (TDNN-Fs) in hybrid DNN/HMM acoustic models. The findings indicated that this architecture outperformed LSTM-based networks in terms of WER. Similar conclusions were drawn in [7] for the Somali language dataset.

A number of papers addressing languages of India has shown effectiveness of TDNNs in tasks related with low-resource ASRs. For example, the authors of [8] presented research of the application of TDNNs, comparing them with bi-directional residual memory networks (BRMN) and bi-directional LSTM. They reported WER of 13.92%, 14.71%, and 14.06% for Tamil, Telugu, and Gujarati, respectively, using the TDNN

and BRMN systems. The authors employed a Kneser-Ney 3-g LM in their study. The introduction of low-rank TDNN with skip connections resulted in an improvement of 0.6–1.1% over the baseline TDNN.

The paper [9] explored the phonetic characteristics relevant to enhancing ASR performance in low-resource Indian languages. They proposed a multilingual TDNN system based on phonetic information. The researchers used a speech corpus provided by Microsoft to construct a system for Gujarati, which exhibited a gradual reduction in WER from GMM (16.95%) to DNN (14.38%) and further to TDNN (12.7%) systems.

Language modeling for low-resource languages is typically performed by n-gram models and recurrent neural network (RNN) based models, with n-gram being applied at the decoding stage, and RNN-based model being applied at the N-best or lattice rescoring stage. For example, this approach was used in [10] for the Sesotho and Zulu languages. The advantage of RNN-based LMs is that they can store the whole context preceding the given word in contrast to feed-forward NNs and n-grams, which store a context of restricted length. It was shown in a range of works, that these types of models have lower perplexity and allows achieving lower WER [11, 12].

Phonemic vocabulary of an ASR system is usually developed automatically by applying some rules converting a sequence of graphemes (letters) to a sequence of phonemic symbols which represent the sounds of speech. When developing ASR systems for Balto-Finnic languages, such as Estonian and Finnish, it is important to consider such features of these languages, as phoneme quantity distinctions. The next section provides the reader with a notion of different approaches to phoneme quantity modeling in ASR for Balto-Finnic languages, focusing on the Finnish and Estonian languages as illustrative examples.

2.2 Approaches to Phoneme Duration Modeling

In Balto-Finnic languages both vowels and consonants exhibit short, long, and overlong (Estonian) quantity degrees [13]. Often these languages are referred to as "quantity languages" due to a significant role of phoneme quantity degrees (as well as other prosodic features like stress and tone). For instance, the variation in the realization of the vowel /a/ as short, long, or overlong in Estonian can result in different meanings for words such as *kalu* ('fish', partitive plural), *kaalu* ('weight', genitive singular), and *kaa:lu* ('weight', partitive singular).

Duration functions in a tool of encoding linguistic information in quantity languages. While some languages, including English, use duration primarily for prosodic purposes such as stress and boundary signaling, quantity languages utilize duration to distinguish between lexical units (see the example above). Studies on various quantity languages have shown that the durational ratios between short and long phonemes remain relatively stable across different articulation rates, indicating their perceptual significance [14]. Absolute durations alone may not be sufficient to convey the quantity distinction, but rather, durational ratios and other acoustic cues contribute to the perception of quantity [15].

When modeling phoneme quantity, researchers typically do not treat different quantity degree representations of the same phoneme type as separate phonological units. Instead, they are represented as one or a sequence of two instances of the

same phoneme [16]. The main reason for this approach is that the determination of long/short and long/overlong quantity degrees goes beyond the characteristics of individual phoneme realizations. It depends on the prosodic variables of neighboring syllables and the over-all syllable/word structure.

Another approach implies treating long/short long/overlong as independent phonemes. For example, in [17] distinctive models for short and long variants of all phones (except /j/) were developed for Estonian. However, the distinction between long and overlong duration is argued to be difficult to model and thus was ignored in acoustic modeling by the authors, being unnecessary in written word forms, as they are not visible in orthography except for a few exceptions.

To model long and short durational ratios, a direct expansion of HMM by including an explicit duration model was used in [18], resulting in what is known as hidden semi-Markov models (HSMMs). Other approaches use forced alignment HMM for the computation of duration features [19, 20]. Consequently, HMM states can be expanded into sub-HMMs that share the same acoustic emission density, allowing for explicit modeling of state durations. This modified model is referred to as the expanded state HMM [21]. Unfortunately, both of these techniques tend to reduce recognition efficiency, as stated in [22, 23].

During the current research the authors investigate the modeling of long sounds by selecting appropriate phoneme set taking into account phoneme duration without modification of HMM framework and topology for Livvi-Karelian ASR.

3 Karelian Text and Speech Corpus

Text and speech corpora are used for training ASR system. The text corpus used within this study is based on the data obtained from publications and journals in Livvi-Karelian. In addition, some texts were imported from the open corpus of Vepsian and Karelian VepKar [24]. Another source for text data were transcripts of audio samples from the training part of speech corpus (see below). The text corpus encompasses diverse styles of speech, such as literary, reportage, and colloquial. A portion of the texts were initially in.pdf format and required semi-automatic text recognition for further processing. All texts were eventually made available in.txt format.

During the preparation of the corpus, the data underwent processing and normalization procedures. This involved segmenting texts into sentences, and converting direct and indirect speech clauses into independent sentences.

Further text modifications were made as well. All texts enclosed in brackets were removed, capital letters were converted to lowercase, and punctuation marks were removed. In earlier Karelian editions the grapheme “ü” can be found, and additional work was made to substitute it with “y”. To ensure the integrity of the textual data, a thorough assessment was conducted to identify duplicate sentences, as the texts were obtained from different sources, so that the duplication of content was highly plausible. The corpus encompassed approximately 5M word occurrences.

One way of speech corpus collection in scenarios involving low-resource languages, established methodologies often involve the active participation of speakers (readers) who read prepared utterances or a coherent text. Another effective approach for collecting speech data entails utilizing freely accessible speech resources. In the present

study, speech data was acquired from radio broadcasts in Livvi-Karelian. A total of 10 broadcasts were used, each broadcast structured in an interview format, featuring a minimum of two speakers (the interviewer and an interviewee). It should be noted that in some broadcasts more than two speakers were present, and interviewers occasionally participated in more than one broadcast. However, no interviewee took part in recording sessions twice. Thus, the recorded speech corpus encompassed 15 speakers, comprising 6 men and 9 women.

The recorded speech data underwent transcription and segmentation (divided into separate statements) procedures conducted by experts in Livvi-Karelian. One significant problem encountered during annotation of texts was simultaneous speech issues, i.e., simultaneous speech from multiple speakers, with interruptions or overlapping. Managing speech overlaps is a complex task, and therefore, phrases containing simultaneous speech of two speakers were excluded from the corpus.

Background noise constituted another factor that hindered the development of the audio corpus. Despite utilizing studio quality recordings, of background noise (music, sounds of turning pages, street noise) were detected. All recordings containing background noise were ultimately removed from the database.

A notable feature of modern Karelian is code-switching [25]. In linguistics, this term generally refers to the spontaneous transition from one language to another. The processing of code-switching in speech recognition demands specialized approaches that were not initially planned for implementation in the system's development. Therefore, all utterances featuring code-switching were excluded from the speech corpus as well.

Proper names present a distinct problem, as they are predominantly borrowed from the Russian language and pronounced according to the Russian phonetic rules. Specifically, stress patterns in names exhibit variability in line with Russian pronunciation. While this problem has yet to be resolved, the most rational solution appears to be compiling a separate dictionary specifically for proper names and transcribing them in accordance with Russian phonetics.

After excluding spoiled segments, the resulting speech corpus amounted to a total duration of more than 3 h (3,819 sentences). The corpus was randomly divided into training and test sets, with 90% of the phrases assigned to the training set and 10% to the test set.

Data augmentation served as an additional tool for expanding the speech data. In this study, augmentation was exclusively applied to the training portion of the speech corpus, utilizing the Sox toolkit [26]. A tempo perturbation augmentation technique was applied to the speech data, the speech rate was varied using a randomly generated coefficient from a uniform distribution ranging between 0.7 and 1.3 for each recording. The augmented speech data was further combined with the authentic training data. As a result, the overall duration of the training data increased from 3 h and 8 min to 6 h and 24 min.

4 Development of a Phonemic Vocabulary

One of the essential prerequisites for developing an automated speech recognition system is the availability of a phonemic transcription dictionary containing words employed by the system. For this purpose, it is necessary to determine a set of phonemes. The

main problem arising when creating phoneme set for Karelian is how to treat long sounds. During the current research several types of phoneme alphabet for Karelian were investigated:

– without distinguishing the long sounds (v1);
– treating the long sounds as independent phonemes (v2);
– long vowels are treated as independent sounds, long consonants are treated as reduplicated of the given sound (v3);
– long vowels, as well as long sonorants and fricative consonants are treated as independent sounds, long plosive consonants are treated as reduplicated phonemes (v4).

It should be noted that in all variants of phoneme set, distinctions were made between stressed and unstressed phonemes, additionally, the back row allophone of the /i/ phoneme was considered as an independent phoneme (/i^/). As for consonants, both palatalized and non-palatalized variants were distinguished. The lists of phonemes used in phoneme sets are presented in Table 1. The transcriptions follow the International Phonetic Alphabet (IPA); additionally, the symbol /!/ indicates word stress, and the symbol /'/ represents consonant palatalization. Symbol /:/ means long sound in these phoneme set variants, which distinguish long phonemes as separate phonemes.

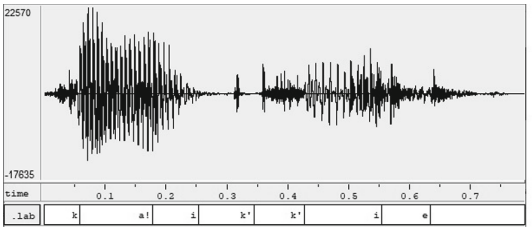
There are two main issues to be noted. Although not all phonemes in the standard Livvi-Karelian have long counterparts, some Livvi-Karelian idioms (mainly, local variants) and borrowings from Russian exhibit long phonemes that are not present in the system of the standard Livvi-Karelian. Due to their infrequent use, it is quite difficult to train acoustic models for such "non-native" long sounds. As a consequence, separate phonemes for these sounds were not introduced (for example, the word *seemejärven* was transcribed as /s' e! m' e j ae r v' e n/). However, when treating long sounds as a sequence of two short phonemes, the "non-native" long phonemes were presented as two separate phonemes (for example, *subbotin* was transcribed to /s u! b b o t'i n/).

The second issue is that in spontaneous speech durational ratios are often reduced, and long sounds may be pronounced as short ones. This is especially true for long Plosive consonants that should be pronounced as two separate sounds, but the second sound is often subject to elision. This is illustrated in Fig. 1 where examples of two realizations of phoneme /k'/ in the word *kaikkie* are shown. In Fig. 1a this sound is realized as a two-sound cluster, one can see repetition of closure and explosion on the waveform. In Fig. 1b, the second sound is omitted and the long phone is realized as a short one. Therefore, when creating phonemic transcriptions for words with long consonants and when treating long sounds as reduplicated ones, two alternative transcriptions were created, namely, a transcription with a reduplicated sound and a transcription with one sound. For example, for word "*kaikkie*" two transcriptions were generated: /k a! i k' k' i e/ and /k a! i k' i e/.

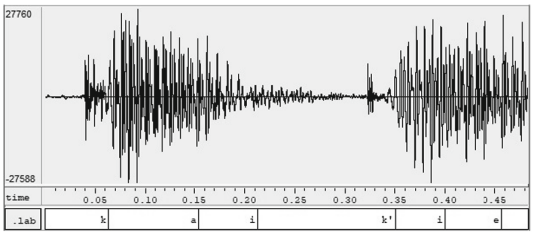
All transcriptions for the vocabulary were created automatically using a software module developed for grapheme-phoneme transformation for Livvi-Karelian. Due to the inherent limitations of automatic recognition techniques for printed Karelian texts, words that occurred only once most often turned out to be incorrectly recognized. Therefore, the dictionary includes all words from the transcripts of the training part of speech corpus and words from other sources that were attested at least twice. The final size of the dictionary was 143.5 thousand words.

Table 1. Types of phoneme sets.

Type of phoneme set	Number of phonemes	Type of phonemes		Phoneme List
v1	53	Vowels	Stressed	/a!/, /o!/, /u!/, /i!/, /i^!/, /e!/, /ae!/, /oe!/, /y!/
			Unstressed	/a/, /o/, /u/, /i/, /i^/, /e/, /ae/, /oe/, /y/
		Consonants	Sonorant	/l/, /l'/, /m/, /m'/, /n/, /n'/, /r/, /r'/, /j/
			Fricative	ch/, /ts/, /h/, /h'/, /f/, /f'/, /s/, /s'/, /sh/, /z/, /z'/, /zh/, /v/, /v'/
			Plosive	/b/, /b'/, /d/, /d'/, /g/, /g'/, /k/, /k'/,/p/, /p'/, /t/, /t'/
v2	90	Vowels	Stressed	v1 + /a:!/, /o:!/, /u:!/, /i:!/, /i^:!/, /ae:!/, /y:!/
			Unstressed	v1 + /a:/, /o:/, /u:/, /i:/, /i^:/, /ae:/, /y:/
		Consonants	Sonorant	v1 + /l:/, /l':/, /m:/, /m':/, /n:/, /n':/, /r:/, /r':/
			Fricative	v1 + /ch:/, /ts:/, /h':/, /s:/, /s':/, /sh:/, /v:/, /v':/
			Plosive	v1 + /d':/, /k:/, /k':/,/p:/, /p':/, /t:/, /t':/
v3	67	Vowels	Stressed	v2
			Unstressed	v2
		Consonants	Sonorant	v1
			Fricative	v1
			Plosive	v1
v4	83	Vowels	Stressed	v2
			Unstressed	v2
		Consonants	Sonorant	v2
			Fricative	v2
			Plosive	v1



(a)



(b)

Fig. 1. Examples of realization of long phoneme /k'/: a) the long sound is pronounced as two sounds; b) the long sound is pronounced as one sound.

In the case of the Karelian language, generating automatic transcriptions represents a relatively straightforward task. This arises from the fixed stress patterns in Karelian, which consistently fall on the initial syllable, while vowel reduction is infrequent. As a result, the automatic transcription process primarily deals with stress localization, identifying dual graphemes as representations of long phonemes, and finding palatalized consonants preceding front vowels.

5 Karelian ASR System

5.1 Acoustic Modeling

Training and testing of a Karelian ASR system was carried out using the Kaldi toolkit [27]. The architecture of the system is shown in Fig. 2.

Hybrid DNN/HMMs acoustic models based on factorized time-delay neural network (TDNN-F) were used. Mel-frequency cepstral coefficients (MFCCs) with additional 100-dimensional i-vector [28] were used as input features to the network.

The core structure of the DNN consisted of three TDNN-F blocks. The initial block was made up of three TDNN-F layers, responsible for processing input vectors (time context of $\{-1, 0, 1\}$). The next block was a single TDNN-F layer (no splicing). The last block comprised ten TDNN-F layers (time context of $\{-3, 0, 3\}$). Each TDNN-F layer had a dimension of 1024, with a bottleneck of 128.

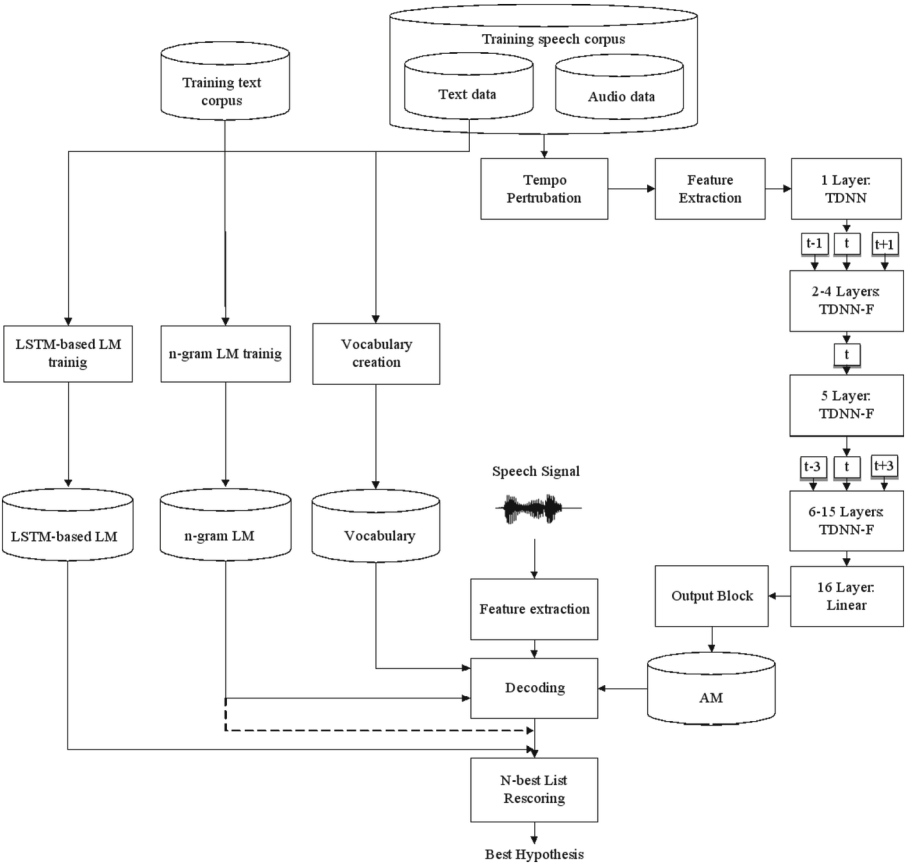


Fig. 2. The Karelian Speech Recognition System.

A Rectified Linear Unit (ReLU) activation function and batch normalization followed each TDNN layers in TDNN blocks. Utilizing skip connections [29], the TDNN layers incorporated the output of each layer (excluding the first layer) by concatenating it with the output of the previous layers. After the TDNN-F layers, a linear layer with a dimension of 256 was employed. The learning rate dynamically adjusted during the training process, starting at 0.0005 and decreasing to 0.00005. The training process was performed in 8 epochs.

5.2 Language Modeling

Both n-gram and LSTM-based LMs were developed, a linear interpolation of these models was made as well. 3-g LM was trained using SRI Language Modeling Toolkit (SRILM) [30]. This model was used at the decoding stage.

LSTM-based LM was trained with the use of TheanoLM toolkit [31]. Experiments were conducted with models with 1, 2 and 3 LSTM layers, the size of LSTM layers was 512. In the models with 2 and 3 LSTM layers dropout at rate 0.5 between LSTM layers

was applied. Optimization criteria was Nesterov Momentum. The initial learning rate was equal to 1. The stopping criteria was "no-improvement", which means that learning rate is halved when validation set perplexity stops improving, and training is stopped when the perplexity does not improve at all with the current learning rate [31]. The maximum number of training epoch was 15.

6 Experiments on Karelian Speech Recognition

The results of experiments on Karelian speech recognition are presented in Table 2. Experiments with different types of phoneme sets, as described above, were conducted. When applying phoneme set v3, two types of phonemic transcriptions were applied: those with alternative pronunciation variants for reduplicated consonants and those with a single pronunciation variant. At the decoding stage 3-g LM was applied, while LSTM-based LM and interpolated models were used at the stage of 500-best list rescoring. In the Table 2 the interpolation coefficient of 0 means that only 3-g LM was used (without 500-best list rescoring). In contrast, interpolation coefficient of 1.0 means that 500-best list rescoring was performed using only LSTM LM.

Table 2. Results of Karelian Speech Recognition in Terms of WER, %

Type of phonemic transcription	Interpolation coefficient for LSTM LM						
	0	0.5	0.6	0.7	0.8	0.9	1.0
v1	27.40	24.78	24.70	24.54	24.78	24.86	25.02
v2	26.29	23.63	23.51	23.55	23.43	23.31	23.51
v3 (without alternative transcriptions for words with long sounds)	25.93	23.43	23.39	23.24	23.24	23.31	23.28
v3 (with alternative transcriptions for words with long sounds)	25.57	23.04	**22.80**	22.88	23.08	23.39	23.67
v4 (with alternative transcriptions for words with long plosive consonants)	25.69	23.67	23.47	23.67	23.87	24.03	24.31

As can be seen from the Table 2, phoneme set with reduplicated consonants (v3) demonstrated better results than this treating long consonants as distinct phonemes (v2). Additionally, the results obtained with this type of phoneme set were better than when using only reduplicated phonemes for plosives. The usage of alternative transcriptions for words with long consonants resulted in additional performance improvement. The best speech recognition results were achieved after rescoring of 500-best list with LSTM LM interpolated with 3-g LM. Application of LSTM-based LM interpolated with n-gram LM with interpolation coefficient of 0.6 for N-best list rescoring resulted in 11% WER relative reduction.

7 Conclusions and Future Work

This paper presents an investigation of different approaches to acoustic modeling for a Livvi-Karelian ASR, focusing on phoneme durations representation issues. Two main approaches were compared within the study: modeling long and short phonemes as separate units vs. representing long phonemes as a sequence of two (or more) short phonemes. The experiments were conducted on a dataset collected by the authors of this paper, and the main metric for evaluation of the results obtained was WER.

The results of experiments have shown, that treating long phonemes as reduplicated units, specifically for plosive consonants, demonstrated superior performance over the approach implying differentiation of long and short phonemes. The usage of alternative transcriptions for words with long consonants further improved the recognition accuracy.

Additionally, different language modeling techniques, including n-gram and LSTM-based models, were investigated. The experiments showed that incorporating LSTM-based language models, especially when interpolated with n-gram models, significantly reduced the WER and improved the overall performance of the developed ASR.

Overall, the idea of using hybrid DNN/HMMs AM with TDNN-Fs combined with LSTM-based LM, demonstrated its effectiveness for processing low-resource languages. The system achieved promising results in WER despite the relatively small amount of training data.

Although the present research has provided positive results in the acoustic and language modeling approaches for low-resource speech recognition, there are several issues to be addressed in future work that can potentially enhance the system's performance:

- Data augmentation: in the experiments, tempo perturbation technique was applied to data augmentation. However, exploring other augmentation techniques, such as spectrogram modification or data generation, could improve the robustness of the developed ASR.
- Incorporating prosodic features: Livvi-Karelian, being a quantity language, relies not only on phoneme durations but also on other prosodic features like stress and tone, to convey different semantical nuances. Future work can explore the embedding of different prosodic models into the current system to process Livvi-Karelian speech more accurately. Additionally, using more advanced techniques, such as hidden semi-Markov models, may result in better representation of phoneme durations and improvement of recognition accuracy.
- Knowledge transfer from other (Balto-Finnic) languages: the techniques and approaches used in this study can be enhanced through models developed for other languages sharing similar phonetic and prosodic characteristics. Investigating the applicability of the data from other (Balto-Finnic) languages, viz. Languages with quantity distinctions as well as the usage of pre-trained multilingual model can contribute to the developed system.

By addressing these issues in future works, the authors of this paper are going to contribute to the development of robust and accurate ASRs for low-resource languages.

Acknowledgements. This research was funded by the Russian Science Foundation, grant number № 22-21-00843.

References

1. Wang, D., Wang, X., Lv, S.: An overview of end-to-end automatic speech recognition. Symmetry **11**(8), 1018 (2019)
2. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4945–4949. The Institute of Electrical and Electronics Engineers (2016)
3. Hori, T., Watanabe, S., Zhang, Y., Chan, W.: Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech), pp. 949–953. International Speech Communication Association (2017)
4. Sun, X., Yang, Q., Liu, S., Yuan, X.: Improving low-resource speech recognition based on improved NN-HMM structures. IEEE Access **8**, 73005–73014 (2020)
5. Karunathilaka, H., Welgama, V., Nadungodage, T., Weerasinghe, R.: Low-resource Sinhala speech recognition using deep learning. In: Proceedings of 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 196–201. The Institute of Electrical and Electronics Engineers (2020)
6. Gupta, V., Boulianne, G.: Progress in multilingual speech recognition for low resource languages Kurmanji Kurdish, Cree and Inuktut. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC), pp. 6420–6428. European Language Resources Association (2022)
7. Biswas, A., Menon, R., van der Westhuizen, E., Niesler, Th.: Improved low-resource Somali speech recognition by semi-supervised acoustic and language model training. In: Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech), pp. 3008–3012. International Speech Communication Association (2019)
8. Pulugundla, B., et al.: BUT system for low resource Indian language ASR. In: Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech), pp. 3182–3186. International Speech Communication Association (2019)
9. Fathima, N., Patel, T., Mahima, C., Iyengar, A.: TDNN-based multilingual speech recognition system for low resource Indian languages. In: Proceedings of 19th Annual Conference of the International Speech Communication Association (Interspeech), pp. 3197–3201. International Speech Communication Association (2018)
10. Wills, S., Uys, P., van Heerden, C.J., Barnard, E.: Language modeling for speech analytics in under-resourced languages. In: Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech), pp. 4941–4945. International Speech Communication Association (2020)
11. Sundermeyer, M., Ney, H., Schlüter, R.: From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(3), 517–529 (2015)
12. Kipyatkova, I.: LSTM-based language models for very large vocabulary continuous Russian speech recognition system. In: Salah, A.A., Karpov, A., Potapova, R. (eds.) SPECOM 2019. LNCS (LNAI), vol. 11658, pp. 219–226. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26061-3_23
13. Metslang, H.: North and standard estonian. In: Bakró-Nagy, M., Laakso, J., Skribnik, E. (eds.) The Oxford Guide to the Uralic Languages, pp. 350–366. Oxford Academic, Oxford (2022)
14. Nakai, S., Kunnari, S., Turk, A., Suomi, K., Ylitalo, R.: Utterance-final lengthening and quantity in Northern Finnish. J. Phon. **37**(1), 29–45 (2009)
15. Traunmüller, H., Krull, D.: The effect of local speaking rate on the perception of quantity in Estonian. Phonetica **60**, 187–207 (2003)

16. Alumäe, T., Vohandu, L.: Limited-vocabulary Estonian continuous speech recognition system using Hidden Markov models. Informatica **15**(3), 303–314 (2004)
17. Alumäe, T.: Recent improvements in Estonian LVCSR. In: Proceedings of 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014), pp. 118–123. European Language Resources association (2014)
18. Kermanshahi, M.A., Homayounpour, M.M.: Improving phoneme sequence recognition using phoneme duration information in DNN-HSMM. J. Artif. Intell. Data Min. **7**(1), 137–147 (2019)
19. Qin, Y., Lee, T., Kong, A.P.H., Law, S.P.: Towards automatic assessment of aphasia speech using automatic speech recognition techniques. In: Proceedings of 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 1–4. The Institute of Electrical and Electronics Engineers (2016)
20. Rosenfelder, I., et al.: FAVE (Forced Alignment and Vowel Extraction) Suite Version 1.1.3. Software. https://doi.org/10.5281/zenodo.9846. Accessed 13 July 2023
21. Johnson, M.: Capacity and complexity of HMM duration modeling techniques. IEEE Signal Process. Lett. **12**(5), 407–410 (2005)
22. Pylkkönen, J.: Phone duration modeling techniques in continuous speech recognition. Master's thesis, Helsinki University of Technology (2004)
23. Pylkkönen, J., Kurimo, M.: Using phone durations in finnish large vocabulary continuous speech recognition. In: Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG), pp. 324–327. The Institute of Electrical and Electronics Engineers (2005)
24. VEPKAR. http://dictorpus.krc.karelia.ru/en. Accessed 13 July 2023
25. Kovaleva, S.V., Rodionova, A.P.: Traditional and Innovative in the Vocabulary and Grammar of Karelian (Based on a Socio-Linguistic Research). KarNC RAN Publ., Petrozavodsk (2011). (in Russian)
26. Sox Toolkit. http://sox.sourceforge.net/sox.html. Accessed 13 July 2023
27. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 1–4. Institute of Electrical and Electronics Engineers (2011)
28. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 55–59. Institute of Electrical and Electronics Engineers (2013)
29. Povey, D., et al.: Semi-orthogonal low-rank matrix factorization for deep neural networks. In: Proceedings of 19th Annual Conference of the International Speech Communication Association (Interspeech), pp. 3743–3747. International Speech Communication Association (2018)
30. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: update and outlook. In: Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), p. 5. Institute of Electrical and Electronics Engineers (2011)
31. Enarvi, S., Kurimo, M.: TheanoLM – an extensible toolkit for neural network language modeling. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech), pp. 3052–3056. International Speech Communication Association (2016)