Training with prosodic information could improve speech-to-text systems

Stanisław Frejlak
s.frejlak@student.uw.edu.pl
University of Warsaw
Supervisor: Prof. Ntalampiras Stavros
University of Milan

***Abstract*—State-of-the-art Automatic Speech Recognition (ASR) systems use graphemes or sub-words as modeling units. They accept raw audio signal as an input, and they output text. However, written forms of natural languages lack certain types of information present in the spoken form, one example being lexical stress. It is an important research question whether including such information in training could improve a model performance. In this work, I use a wav2vec 2.0 model pre-trained on multi-lingual unlabeled data, and I fine-tune it on a set of utterances in Russian language. I find out that a system trained against text with added lexical stress markings outperforms a system trained against raw text.**

Keywords: automatic speech recognition, lexical stress, modeling units, wav2vec2

1. Introduction

In natural languages, there is no one-to-one mapping between *graphemes* - signs used in the written language, and *phonemes* - the smallest units of the language spoken form. Moreover, phonemes tend to have multiple *allophones* - different ways of pronouncing them. This has been one of the reasons for which an automatic speech-to-text transcription has been a difficult task in the field of machine learning.

Early approaches to the task of speech-to-text transcription were dividing the process into a few phases [1]. Engineers needed a big linguistic knowledge to transcribe a text into a sequence of phones. As the phoneme pronunciation is context-dependent, the systems used sequences of a few consecutive phones as modeling units. When a system, e.g. a Hidden Markov Model, or a neural network, achieves a satisfactory accuracy in predicting the phones, another task is to decode the original text from these predictions. This is a source of a new difficulty, as for most languages, the mapping from sequences of graphemes to sequences of phones is not bijective. Reversing this mapping requires having a language model, either hand-crafted, or trained.

Recent work shows that all this tedious linguistic work is not necessary when training attention-based models on big amount of data. Radford et al. [2] published the Whisper model, trained on 680K hours of audio data, which closes the gap to human performance in speech recognition. A neural network is able to master phones recognition alongside with learning a language model, so it can directly output a written form of an utterance.

While using written form of a language brings good results, it is still valid to ask whether incorporating some additional information about the words pronunciation could improve model results. A written form of a natural language does not include a lot of information present in speech, such as phonological or prosodic features.

A lexical stress is an example of such information which is highly important in some languages. For example, in English or Russian, there exist homographs whose pronunciation differ by the position of lexical stress. L2 speakers often make mistakes by placing the stress on the incorrect syllable. It is of interest to create ASR models which would include the stress marking in their outputs, and there is a good body of research concerning this topic. In some languages, the position of lexical stress in a word strongly influences the pronunciation of many of the word's phonemes. Therefore, one might wonder if including the information about lexical stress might help ASR models also in the task of producing transcriptions without stress annotations.

Baevski et al. [3] presented a wav2vec 2.0 model pre-trained in a fully unsupervised manner on 56K hours of utterances in 53 languages. The model can be later fine-tuned on a very small amount of labeled data in any language to achieve a low Word Error Rate (WER). As the pre-trained model is not biased towards any type of labels, it is feasible to fine-tune using various forms of utterance representations.

In this work, I compare the performance of the wav2vec 2.0 model fine-tuned in two different experimental settings. In both of them, the model sees 20 hours of utterances in Russian language as training data. For one model, the labels are simple transcriptions of sentences, and for the other, the transcriptions are enriched with the lexical stress markings and some basic phonological features.

2. Related work

Choice of modeling units has been one of important issues in the field of automatic speech recognition [1][4]. In the classical approaches to Automatic Speech Recognition, phoneme-based models outperformed grapheme-based models. However, the former approach poses problems connected with transcribing text into a sequence of phones,

and then the other way around, predicting text from a model output representing a sequence of phones. This proves to be very difficult even for languages with seemingly very close correspondence between graphemes and phonemes. For example, Mirilovič et al. [5] presented an analysis of this problem for Slovak language.

Irie et al. [4] showed that contemporary attention-based encoder-decoder models perform better when using character-based modeling units. In some languages, such as English, the correspondence between orthography and pronunciation is not very clear, but in this case, using sub-words as modeling units appears to be the best choice. The authors conclude that such a model, when trained on a large amount of data, jointly learns the acoustic model, pronunciation model, and language model within a single neural network.

A striking evidence of strength of this approach is shown by Zhou et al. [6] in the context of Chinese language. Unlike the Latin alphabet, Chinese writing system consists of thousands of logograms representing words or sub-words. Among the characters, there is a huge amount of homonyms. For each syllable present in Chinese language, there are many characters which are pronounced this way. For such a writing system, it seems especially difficult to train a model using grapheme-based representations of utterances. However, as shown in [6], models using Transformer architecture trained with grapheme-based modeling units achieve better results than when the modeling units are phoneme-based.

The success of grapheme-based models did not lead to the discontinuation of research on predicting phonetic representations of utterances [7][8]. These studies are especially needed for building ASR models for low-resource languages. To this goal, Baevski et al. [9] and Xu et al. [10] utilize multilingually pre-trained wav2vec 2.0 model for producing phonetic transcriptions for unseen languages. In [10] the authors incorporate the linguistic knowledge by mapping phonological features of phonemes from languages seen during pre-training to languages not seen in that phase which the model is fine-tuned on.

Most notably for this paper's topic, Kubo and Bacchiani [11] show that attention-based ASR models benefit from joint training on tasks of predicting text representation and phonetic representation of utterances.

Automatic recognition of lexical stress in utterances has been studied since 60s [12]. In 2020 Korzekwa et al. [13] succeeded in exploiting the attention mechanism to build a stress recognizer without the need for manual feature extraction from the audio signal. Frejlak et al. [14] used wav2vec 2.0 pre-trained model for the task of stressed syllables recognition. The study showed that extraction of such prosodic features like lexical stress can be conducted with an ASR model pre-trained in an unsupervised manner.

To my best knowledge, no work to date has checked whether including the information on lexical stress in end-to-end training of attention-based ASR model could improve its performance. However, the results of [11] might suggest that such improvement could be observed.

3. Problem

In all languages, phones can be grouped by a number of phonological and prosodic features. In Russian, the phonological features of consonants include voicing, palatalization, place of articulation, and manner of articulation. Vowels could be differentiated by phonological features such as height, backness, and palatalization, as well as the prosodic feature of stress.

Including all this information in training of an ASR model theoretically might be helpful. However, it is a very complicated task since letters of Russian alphabet are pronounced differently, depending on many contextual factors. For example, letters representing voiced consonants are pronounced voiceless at the end of a word. There are more rules about the exact pronunciation of letters, some of which are complicated and have many exceptions.

The other problem is finding a way to incorporate this information in the training pipeline. As each sound can be characterized by a few features, the network outputs should be not a one-dimensional vector of probabilities for all used token but rather a set of vectors, or a matrix. A correct design of such an architecture is not an easy task, and it cannot be performed using out-of-the-box functionalities of ASR systems published to date.

The last problem is an automatic transcription of outputs containing the information about phonological and prosodic features to text. As mentioned before, this task is difficult and requires a good language model. Alternatively, as suggested by Kubo and Bacchiani [11], the neural network could have two outputs: text and phonetic transcription. However, if one intends to use the model e.g. for automatic stress annotations, then combining these two outputs might be still challenging.

In this work, I make sure that utterance transcriptions could be obtained from ASR model predictions without a need of training any additional model. I preprocess ground-truth transcriptions in a way which can be trivially reversed to the original text. I focus on two features: palatalization and lexical stress.

Palatalization is an important distinctive feature of phones in Slavic languages. Often, it is not just a feature of a single phone, but rather of both a consonant and a vowel which appears after it. Different Slavic languages denote this feature in different ways. For example, in Polish language, palatalization is denoted by an additional letter "i" placed between a consonant and a vowel. Palatalized consonant not followed by vowels are marked with diacritics. On the other hand, in Russian language there is a whole set of different letters denoting iotified vowels. A soft sign is written after palatalized consonants not followed by vowels.

As the palatalization process arguably can be heard between the consonant and a vowel, I preprocess the transcriptions by replacing iotified vowels by a soft sign and the vowel's non-iotified counterpart. This operation does not reflect all the rules concerning palatalization in Russian language but I theorize that it can still be helpful for the model as it decreases the number of modeling units and exposes one distinctive feature of phones.

Lexical stress has a big significance in Russian language. Each basic form of a word has a fixed syllable on which stress should be put. Words are classified into many categories according to the way in which stress position changes in their inflected forms.

Not only does a lexical stress modify pronunciation of the stressed vowel, but also it affects the way in which other vowels in the word are pronounced. There are many rules about the reduced pronunciation of vowels depending on the placement relative to the stressed vowel. Moreover, the exact pronunciation might depend on such factors as the speed of speech, or a dialect of the speaker.

Transcribing non-stressed vowels into phones is a cumbersome task and it cannot be easily reversed, as different vowels share their reduced forms. Therefore, in the transcriptions, I leave the non-stressed vowels the way they are written but I introduce new tokens for the stressed vowels.

I theorize that the two described text modifications connected with palatalization and lexical stress might improve results of an ASR system.

4. Experimental Set-up

I fine-tune two models, both based on a pre-trained XLSR-Wav2Vec2 model [15]. The fine-tuning is done on Google Colab's standard GPU, using 20 hours of audio data. Due to hardware and resources limitations, the models are not expected to achieve state-of-the-art performance. However, the comparison of the two model's results anyway provides a clue in the discussion about the choice of modeling units for ASR systems.

4.1. Training data

Similarly as in my previous work [14], I train the model on short speech recordings taken from the Multimodal subcorpus of the Russian National Corpus [16]. The recordings were collected using RNC API, by queries with a few thousand most frequent Russian words. I cap the length of the used recordings by 15 seconds, and discard those with more than one speaker. Training data consists of 19,5 hours of data and the evaluation data - of 1,5 hours.

Every example in the Multimodal subcorpus contains not only a recording of speech but also a stress-annotated utterance transcription. The transcriptions are of high-quality as they were prepared by a team of linguists. I use these transcriptions to prepare training labels.

4.2. Training labels

In one experimental setting, labels for the recordings are simply their transcriptions, with all punctuation and stress markings removed.

In the other experimental setting, I preprocess the transcriptions to include the information about palatalization and lexical stress. I replace tokens corresponding to iotified vowels with their non-iotified counterparts and put a soft sign (ь) before them. In case of a iotified vowel starting a word or being written after a soft or hard sign, I put letter й in front of it, since in this case, the voiced palatal approximant consonant is pronounced, and it is generally denoted in Russian by letter й. I also get rid of hard signs from the text.

Because of Russian orthographic rules no two words could be mapped into the same text by this preprocessing. Reverse operations include putting a hard sign between consonant and letter й and replacing vowels preceded by letter й or a soft sign with their iotified counterparts. This way all preprocessed words are mapped back to their original forms with a few exceptions of borrowed words like "йогурт" which would be mapped back to "ёгурт".

As for the lexical stress, I introduce new tokens corresponding to stressed vowels. The Russian National Corpus includes also markings for the so-called secondary stress. I denote the vowels with secondary stress in the same way as vowels with the primary stress. Reversing this operation is done by replacing these new tokens with their not stressed counterparts.

4.3. Training set-up

To train the two models, I use the HuggingSound [17] toolkit which is built upon the HuggingFace library. I use batch size of 24, with instantaneous batch size of 12 and two gradient accumulation steps, in order to reduce the GPU RAM usage. For the same reason, I use gradient checkpointing and `FP16` precision computations. In other experiments, I used 8-bit Adam optimizer, but then the training did not converge.

I run the fine-tuning for four epochs. I use learning rate of $1e-4$, as in my experiments it brought better results than the default learning rate of $3e-4$ proposed in the Hugging Sound library. The learning is scheduled with a number of warm-up steps corresponding to the length of the first epoch.

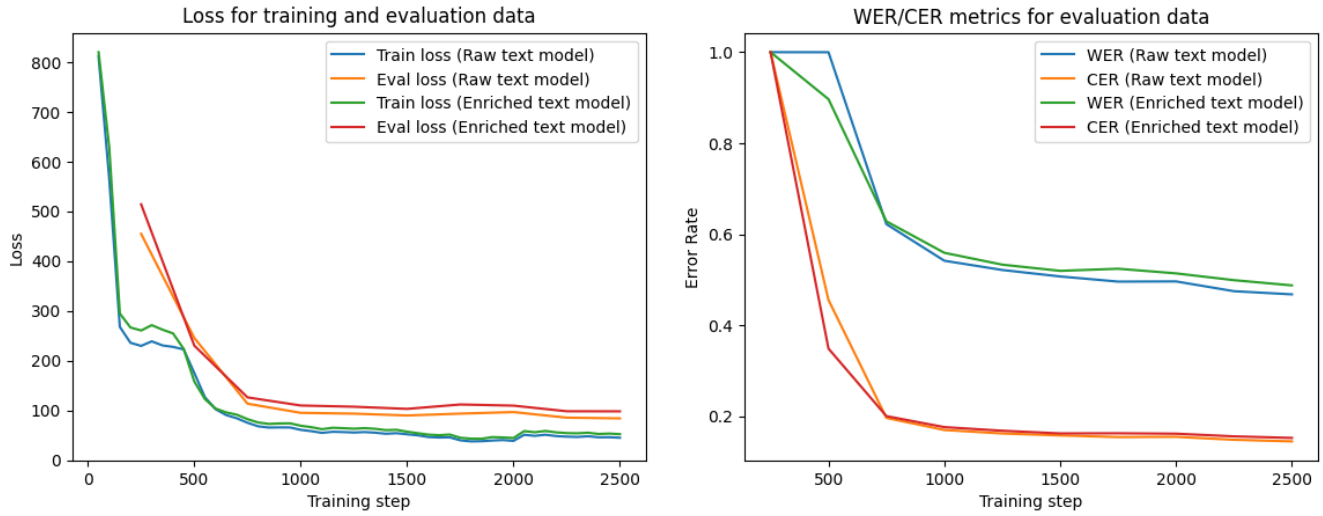
In multiple experiments with different settings, I observed a consistent tendency that by the end of the fourth epoch the loss function reaches a plateau after which it starts to grow. I resume the training from a checkpoint at the start of the plateau (training step 1000, epoch 3.89) using a learning rate of $3e-5$. This phase of the training also uses learning rate warm-up of the length of one epoch, and decay in the sixth and last epoch. In total, there are 2542 training steps (10 epochs).

My code is available on GitHub [18].

5. Results and discussion

5.1. Results

The two models are trained with different tokens and the metrics during the course of training are counted for both of them with their respective token sets. As shown in figure 1 the training and evaluation loss decrease with time in a very similar way for both models. The same goes for



(a) Loss plot

(b) Metrics plot

Figure 1: Evolution of loss and metrics for both experimental set-ups look very similar. The error rates in plot 1b are not directly comparable because they are computed for different token sets.

WER and CER metrics. One can see that the error rates for the model trained with preprocessed transcriptions are a bit higher. However, these metrics are computed using transcriptions which include stress markings which means that the prediction is a harder task.

The most important result comes from comparison shown in table 1. When the transcriptions predicted by the model trained on preprocessed text are processed back to raw text, this model achieves better scores both for WER and CER metrics. The difference in WER metric is significant.

	WER	CER
`Raw text model`	0.4727	0.1462
`Enriched text model`	**0.3857**	**0.1398**

Table 1: Comparison of both model's error rate metrics counted for raw transcriptions

5.2. Discussion

Newest advances in attention-based ASR research show that state-of-the-art results could be achieved with using raw text as labels. However, utilizing linguistic knowledge might still be helpful when designing ASR systems. As shown by Kubo and Bacchiani [11], using prediction of phonetic transcription as an auxiliary task improves model performance. Results of my work are in line with those findings and suggest that predicting prosodic features of utterances might also help in the main task of utterance transcription.

The models trained in this work used only 20 hours of audio data which is not a big amount compared to datasets used in contemporary works, also for Russian language. However, for many languages in the world there is even less audio data available so training in such setting is also an important part of research.

Analysis of mistakes in predictions show that the used resources were not enough for the networks to learn a full language model. For example, in the transcribed sentences one can find non-existing words like "повот" in place of a word "повод". This mistake corresponds to the articulation of the word, as letter д is replaced by its voiceless counterpart т and in speech voiced consonants at the end of a word are pronounced voiceless.

5.3. Future work

Further experiments might establish if the results obtained in this work extend to models trained on larger datasets. With more data, it is expected that a model might learn the exact orthography of most of the words present in a language. However, adding the information about lexical stress might be still beneficial, especially for languages with complicated rules concerning moving of the stress in inflected forms of a word.

In the context of low-resource languages, the study suggests that manual extraction of language-specific distinctive features of phones might help in creating ASR systems. Obviously, using raw text as labels in training is the cheapest solution. However, for many languages it might be possible to easily obtain some additional information about utterances given their transcriptions.

References

[1] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," *Proc. Eurospeech*, 04 2009.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, and P. Nguyen, "Model unit exploration for sequence-to-sequence speech recognition," *CoRR*, vol. abs/1902.01955, 2019. [Online]. Available: http://arxiv.org/abs/1902.01955

[5] M. Mirilovič, J. Juhár, and A. Čižmár, "Comparison of grapheme and phoneme based acoustic modeling in lvcsr task in slovak," in *Multimodal Signals: Cognitive and Algorithmic Issues*, A. Esposito, A. Hussain, M. Marinaro, and R. Martone, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 242–247.

[6] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," 2018. [Online]. Available: https://arxiv.org/abs/1805.06239

[7] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," *CoRR*, vol. abs/2002.11800, 2020. [Online]. Available: https://arxiv.org/abs/2002.11800

[8] C. Leong and D. Whitenack, "Phone-ing it in: Towards flexible multi-modal language model training by phonetic representations of data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.364

[9] A. Baevski, W. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *CoRR*, vol. abs/2105.11084, 2021. [Online]. Available: https://arxiv.org/abs/2105.11084

[10] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *ArXiv*, vol. abs/2109.11680, 2021.

[11] Y. Kubo and M. Bacchiani, "Joint phoneme-grapheme model for end-to-end speech recognition," in *Proc. ICASSP 2020*, 2020.

[12] P. Lieberman, "Some acoustic correlates of word stress in american english," *The Journal of the Acoustical Society of America*, vol. 32, no. 4, pp. 451–454, 1960.

[13] D. Korzekwa, R. Barra-Chicote, S. Zaporowski, G. Beringer, J. Lorenzo-Trueba, A. Serafinowicz, J. Droppo, T. Drugman, and B. Kostek, "Detection of lexical stress errors in non-native (l2) english with data augmentation and attention," *arXiv preprint arXiv:2012.14788*, 2020.

[14] S. Frejlak, J. Rutkowski, and J. Bednarz, "Detecting lexical stress for Russian language," 2022.

[15] "Xlsr-wav2vec2." [Online]. Available: https://huggingface.co/docs/transformers/model_doc/xlsr_wav2vec2

[16] E. Grishina, "Russian national corpus," 2006. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/92_pdf.pdf

[17] "Hugging sound." [Online]. Available: https://github.com/jonatasgrosman/huggingsound

[18] "Stress detection." [Online]. Available: https://github.com/siasio/StressDetection