ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему: Что модель Wav2Vec-2.0 выучивает в процессе изучения нового языка

Тема на английском: *What Wav2Vec-2.0 Learns While Fine-tuning to New Language*

Студентка 2 курса группы
БКЛ-201
Шерман Ксения Валерьевна

Научный руководитель
Сериков Олег Алексеевич
Приглашённый преподаватель

Москва, 2023 г.

Contents

1. Introduction

Transformer models have drawn a lot of attention since their appearance as they show success in solving various tasks such as computer vision, image generation, speech reproduction and recognition. The most popular examples of such models are voice assistants like Yandex Alisa and chatbots like ChatGPT. The actions inside the model allow it to recognize the request, formulate an answer and produce it. Nevertheless, the perfect, i.e., state-of-the-art, results are followed by the poor interpretability of the models or the black box problem. It is unknown how AI models learn and what they learn while training. In other words, what makes these models so powerful.

The main reason of uninterpretability is that the structure of transformer models is quite complex. It consists of several connected hidden layers each of which has a certain number of neurons, or parameters. These parameters change during the process of training, adjusting to the data. However, it is unclear why the parameters change the way they do. Moreover, another question is whether there is any knowledge about the structure of the data in the model, for example, is it able to learn linguistic properties of the data.

There are two types of a research: analyzing the result of the model's training and analyzing the process itself. The former type tries to answer the question: what information is learned and what the representations mean. The latter tries to define what changes are made in the internals of the model and how vector space is created. As language models are usually trained to solve a new problem for a few epochs after the main training is done, the second type of research is of great interest. This main training is called pre-training and generally is followed by fine-tuning.

Though a number of articles are written about the structure of text models, especially BERT [Devlin et al. 2018], a little research is made in the field of acoustic speech recognition (ASR), that is, when a speech recording is translated into text. The ASR problem is solved by fine-tuning a pre-trained speech model to predict the text transcription. There are several known models which are trained to solve this task: Wav2Vec-2.0 [Baevski et al. 2020], Whisper [Radford et al. 2022], Hubert [Hsu et al. 2021]. In this work the Wav2Vec-2.0 model is considered. We attempt to look at the internals of the model pre-trained on English data and describe the process of its fine-tuning with Russian language recordings to study whether the model is able to learn a new language. The main tasks are:

1. Fine-tune the model and run tests obtaining information from each layer of the model;
2. Define the research methods for model's interpretation and structure analysis;
3. Define the sample of sounds for a research based on phonetics of English and Russian languages.

The model cuts a recording into several segments and produces a vector for each segment at each layer. The audio representation is a list of segments vectors. The last layer of the model tries to match each vector with a label of a heard sound. The main idea is that these vectors change during the fine-tuning but it is unknown how exactly. Therefore, we stated such hypotheses about the process of fine-tuning:

1. The model can separate English sounds from the Russian ones. In other words, the model encodes the information about the language of a sound.
2. The model mixes the knowledge about English and Russian sounds at the end of the training. This hypothesis is based on the assumption that before training the model does not know Russian language thus it distinguishes Russian sounds from the English ones. However, during the training it may learn the similarities between two languages and mix the sounds.
3. The representations of Russian sounds change more than the representations for the English language. The idea behind this hypothesis is that the wav2vec model is pre-trained on English data and should not relearn it while the Russian audio are heard for the first time.
4. First layers of the model change less than the last ones during the fine-tuning. The first layers of models are usually responsible for the general knowledge while the last ones are trained to solve a specific task. Thus, they tend to change more.
5. The representational space of the model changes: the model creates clusters of similar sounds which move apart from each other. During the fine-tuning the model better learns the differences between sounds, hence, it makes the vectors of distinct sounds less similar.

Additionally, we propose a few hypotheses about the model's phonetic knowledge.

1. The model perceives Russian voiced plosive consonants as more similar to English voiced consonants between vowels than to English voiced consonants at other positions in a word at the initial stages of learning.
2. The model perceives Russian voiceless plosive consonants as more similar to English voiced consonants at the beginning of a word than to English voiceless consonants at the initial stages of learning.

The both of these hypotheses are based on the voice onset time (VOT) of the plosive consonants like [d], [t], [b], [p] etc. The first hypothesis is based on the claim that Russian voiced stops have VOT similar to English consonants in the position between vowels. The idea behind the second hypothesis is that the VOT of English consonants at the beginning of the word is close to VOT of Russian voiceless stops. It is expected that the model perceives a sound from a new language as similar to the sounds it knows. More information about VOT is in §2.3.

The study was conducted using the Python programming language. The code notebooks are in the GitHub repository: https://github.com/ShermanKsenia/fine-tuning-probing.

2. Literature review

2.1. Wav2Vec-2.0

Wav2Vec-2.0 [Baevski et al. 2020] was introduced in 2020 and belongs to the class of transformers. It has three blocks which process the data: feature extractor, context network, and quantization module. First block is a multi-layer convolutional neural network and it is responsible for raw audio transformation to a hidden state representation by dividing a waveform into segments of equal length to make a discrete input from continuous one. Next, the context network creates new representations obtaining information from all segments of an audio. They are then used for training. While learning the model solves a contrastive task which is based on the prediction of masked audio segments. The goal is to find the right quantized representation corresponding to the masked segment from the set of distractions. Quantized representations are computed in the quantization module after an audio is segmented, which makes it possible to obtain a finite set of speech representations.

In this work we use wav2vec2-base model which was pre-trained on 960 hours of English speech audio from Librispeech dataset [Panayotov et al. 2015]. It contains 12 hidden layers with 768-dimension space and 8 attention heads. It is used as a feature extractor and requires a new classification layer to be made to solve an ASR task. The model is downloaded from Hugging Face repository.

2.2 Language Models Interpretation

One of the most popular methods to study models is probing which is «a classification problem that focuses on simple linguistic properties of sentences» (Conneau 2018: 2126). It requires creating an additional classifier, for instance, logistic regression, which tries to learn some linguistic property, like part of speech, from the model representations. If a classifier succeeds to learn it, that means that the model encodes this information in a vector. The success is measured by accuracy score: the higher the score, the better information is encoded in embeddings. Thus, it is show that the BERT keeps the information about a part of speech of a word [Tenney et al. 2019]. Moreover, this method allows to compare hidden layers by their ability to keep information about a linguistic property. It is used in article "Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings" [Wiedemann et al. 2018]. It is shown that high layers of the model encode knowledge about the context better than low layers.

Probing is used for studying Wav2Vec-2.0 and other models in [Shah et al. 2021]. Three ASR neural networks are probed on different tasks like speaker identification or prediction of sound properties. They conclude that the models encode information about speaker and phonetics, for instance, neural networks differ consonants and vowels and know about distinctions between plosive and fricative consonants.

Whereas a labeled dataset is required to perform probing, there are methods which can be used without additional dataset. The modification of Representational Similarity Analysis (RSA) [Kriegeskorte et al. 2008] - Representational Stability Analysis (ReStA), - is used to study vector space in BERT, ELMO and GoogleLM [Abnar et al. 2019]. They vary the length of the context applied to the input of the model and compare the representations from first and second layers of the model. As a result, it is shown that first layers in LSTM models are less sensitive to changes in the context length while the first layers in transformers change more than the last ones.

Another method similar to RSA is Canonical Correlation Analysis (CCA). Elena Voita, Rico Sennrich and Ivan Titov use its modification, PWCCA, to compare representations from the models fine-tuned on three tasks: machine translation (MT), language modeling (LM) and masked language modeling (MLM) [Voita et al. 2019]. Consequently, they show that MT and MLM models produce representations which are closer to each other than to LM's output. They explain such a behavior by saying that MT and MLM models have similar structure, for example, they focus on a given token to translate or predict, whereas LM model tries to predict the next token in a sequence. Moreover, they compare trained representations from models with different initializations and conclude that PWCCA better captures distinctions between the models fine-tuned on distinct tasks than between models initialized differently. In other words, PWCCA better captures differences in the types of information learned.

The process of fine-tuning is studied as well. Another modification of CCA – SVCCA is used to explore learning dynamics and to evaluate the changes of feature extraction mode [Hao et al. 2020]. Authors conclude that main alterations occur in the middle and last layers of the BERT model while the first layers are stable. They conclude that lower layers are likely to contain more general knowledge whereas task-specific information is encoded in the higher layers.

Fine-tuning process in BERT was studied by Yichu Zhou and Vivek Srikumar [Zhou, Srikumar. 2021a]. They conduct several experiments based on probing and DirectProbe [Zhou, Srikumar. 2021b]. This method analyzes the geometry of the vector space. Thus, they study the clusters where each vector has the same label and measure the number of clusters, distances between clusters and spatial similarity. Yichu Zhou and Vivek Srikumar train a linear SVM to separate clusters and then compute the margin of maximum margin separator. As a result, it is shown that fine-tuning makes the space simpler, namely, decreases the number of clusters meaning that the model learns generalization rules. In addition, fine-tuning pushes the labels away from each other. Moreover, it is stated that high layers remain close to the original representations but are still less similar to the initial representations as expected. This is due to the fact that the model attempts to preserve the pre-trained information encoding new task-specific information at the same time.

2.3 Phonetics

Plosive consonants in languages differ by voice onset time (VOT). It is defined as the interval between the release of a stop consonant and onset of voicing, i.e., vibration of the vocal cords. There are three types of VOT: voicing lead, short voicing lag, and long voicing lag. The difference is depicted in Figure 1.

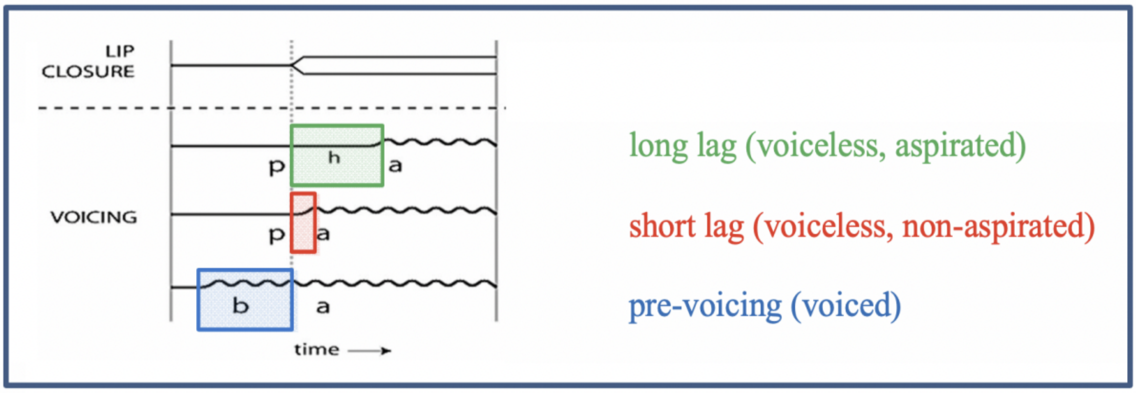
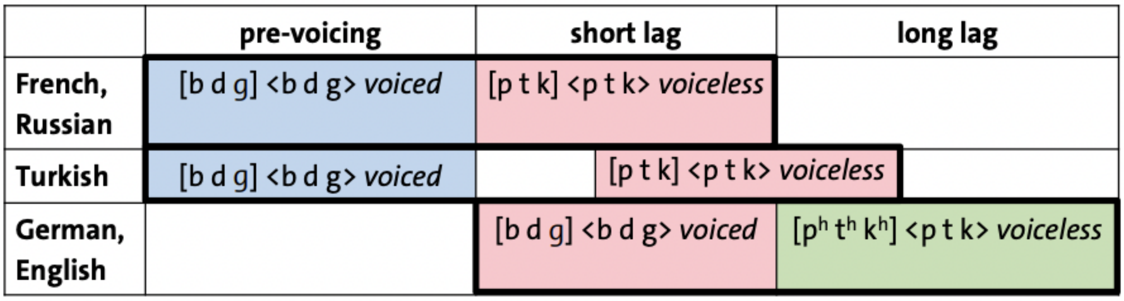

LIP CLOSURE
VOICING
p
h
a
p
a
b
a
time →
long lag (voiceless, aspirated)
short lag (voiceless, non-aspirated)
pre-voicing (voiced)

Figure 1. The picture is from [Dittmers et al. 2017]

By a default, the time of a release of a closure equals 0. Sounds with long lag VOT are called aspirated stops and have a long interval between the moment of closure and a voicing, positive one. Short lag VOT length is approximately 0. The segments with short lag VOT are called unaspirated stops. The last type is defined by the negative VOT meaning that the voicing starts before the moment of closure. Such stops are called prevoiced. The time when the voicing begins varies from one language to another. Similar sounds in distinct languages may differ in their VOT. As it is shown in Figure 2, the list of languages can be made to divide them into separate categories by the VOT of stops.



	pre-voicing	**short lag**	**long lag**
French, Russian	[b d g] <b d g> *voiced*	[p t k] <p t k> *voiceless*	
Turkish	[b d g] <b d g> *voiced*	[p t k] <p t k> *voiceless*	
German, English		[b d g] <b d g> *voiced*	[p^h t^h k^h] <p t k> *voiceless*

Figure 2. The picture is from [Dittmers et al. 2017]

We are interested in Russian and English sounds for the study. Russian and English short lag sounds are voiceless and voiced respectfully. It means that there are such combinations of sounds where voiced English stops are similar to Russian voiceless stops. That is why Ringen and Kulikov [Ringen, Kulikov. 2014] finds that Russian and English languages cannot be compared in terms of “voiced” and “voiceless” sounds as English initial voiced stops do not have pre-voicing (as Russian voiceless stops) while Russian voiced ones do. Moreover, Russian voiceless plosive consonants lack an aspiration compared to English stops. Some studies suggest that voicing might be changed while learning a new language [Dittmers et al. 2017; Zaikovskii, Koffi. 2019]. In both work it is concluded that L2 speakers can acquire the VOT of the second language.

3. Methodology

The crucial goal of the work is to define the changes in the representational space of the model. Each sound corresponds to some vector. We obtain all vectors from each layer and after each epoch of fine-tuning to conduct an analysis. Thus, there are a few methods which are used in this study.

3.1 Fine-tuning and Data

Fine-tuning is done in order to teach the model to solve the task of transferring oral speech to text. Consequently, it is necessary to build a classification layer that correlates the vector and the sound it represents. The output of this layer is called logits. We fine-tune the English based model to predict text in English and Russian languages. The English data is used mainly to ensure that the model does not forget the English language.

There is no dataset with phonetic transcriptions of Russian speech, thus, the work is based on text data, namely, the model predicts a letter rather than a sound. Therefore, it is important to check whether the obtained results are somehow related to phonetics of languages or only letters. We compare vectors of three pairs of “sound-letter”: [Λ] – “o” in a word “корова” (“cow”), [Λ] – “a” in a word “азбука” (“ABC”), [ɔ] – “o” in a word “облако” (“cloud”), - in wav2vec model trained on Russian language. We calculate cosine similarity between each pair of sounds. The similarity of the same sound is higher in the lower layers meaning that first layers encode phonetics better. The closer the audio vectors are to the output of the model, the more they represent the letters.

The hyperparameters for the fine-tuning are: learning rate = 0.0001, weight decay = 0.005, warmup steps = 1000. We trained the model for 18 epochs. After each epoch we saved the model’s parameters to be able to extract audio representations of sound from each layer on each epoch. The results of training are shown in Appendix B.

We fine-tune the model on two datasets for each language: TIMIT [link] with English audio and CommonVoice [link] with recordings in Russian. TIMIT consists of 6300 recordings of 630 English speakers. In CommonVoice there are more than 20000 audio recordings of Russian speech. The metadata for each speaker is provided in both datasets. For fine-tuning we use 2000 recordings from each of the datasets as a train dataset and 500 recordings as a test one. The results of fine-tuning are shown in Appendix B. The predictions of fine-tuned model are in Table 1.

	Prediction	Target
English	she had your dark suit in greasy wash water all year	she had your dark suit in greasy wash water all year
Russian	èkoomičeskoâ politi ka kuby sozdaût ser′eznoâ prepârtstve	èkonomičeskaâ politika kuby sozdaet ser′eznye prepâtstviâ

Table 1. The predictions of fine-tuned model and the target sentences

The initial state of the model, or the state before fine-tuning, is used for comparing the results of training. We use the term *zero epoch* to refer to this state.

3.2 Phonetics of the experiments

We chose the next pairs of sounds: [b]/[p], [d]/[t], [g]/[k]- as they differ in VOT in Russian and English languages. The pronunciation in English depends on the place of a sound in a word. Thus, the English stops [d], [b], [g] in the initial position in a word are pronounced with short lag VOT as Russian voiceless stops. These stops in the middle of a word are mostly pre-voiced as voiced stops in Russian language [Ringen, Kulikov. 2014]. Therefore, we obtain sound representations with the information about the position in a word. In addition to these sounds, we chose the pair of [f]/[v] as these sounds are almost similar in both languages and they should show different result.

For better understanding, we add to labels from Russian data suffix "_ru" and to labels from English data suffix "_en".

For testing the phonetic hypotheses, we compute the distances between vectors of each pair of Russian and English sounds using these formulas for the first and the second hypotheses respectively:

1. Distance between "C_ru" and "C_en" between vowels is compared with distance between "C_ru" and "C_en" at the beginning of a word for C in a set: "b", "d", "g", "v".

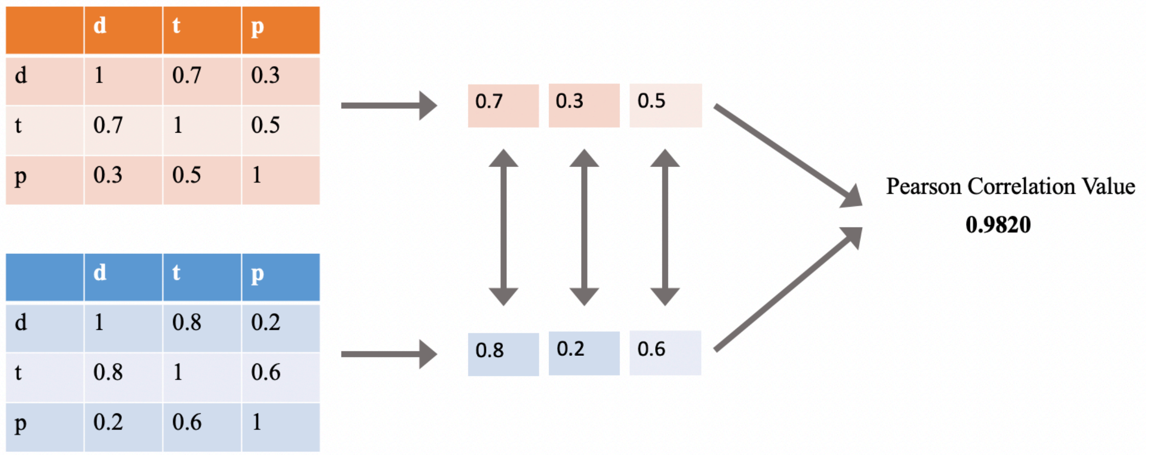
2. Distance between “C_ru” voiceless and “C_en” voiceless at the beginning of a word is compared with distance between “C_ru” voiceless and “C_en” voiced at the beginning of a word for C in a set: “p”/ “b”, “t”/ “d”, “k”/ “g”, “f”/ “v”.

For instance, we compare the distance between “p_ru” and “p_en” at the beginning of a word with the distance between “p_ru” and “b_en” at the beginning of a word for the hypothesis 2.

The number of sounds used for these experiments is written in Appendix C.

3.3 Representational Stability Analysis

Representational Similarity Analysis (RSA) [Kriegeskorte et al. 2008] was introduced to compare neuron activations of a human brain and activations of a model. As for this study, the method allows to measure the similarity between the representations obtained from the model after different epochs. This modification is called Representational Stability Analysis (ReStA). First step is to compute cosine similarity between each pair of sounds in the dataset to create the similarity matrix for each of two representational spaces. For each sound we calculate the average sound vector among all its representations. Second step is to calculate Pearson correlation between two matrices to quantify the similarities across the representational spaces. The idea behind RSA is that the relative configuration of the representational space can stay unchanged but the space itself can be rotated. Thus, this rotation is neglected as it is unimportant. Example of the process is depicted in Figure 3.



	d	t	p
d	1	0.7	0.3
t	0.7	1	0.5
p	0.3	0.5	1

	d	t	p
d	1	0.8	0.2
t	0.8	1	0.6
p	0.2	0.6	1

Figure 3. First, cosine similarity of each pair of sounds is computed in a matrix. Then, the numbers above the diagonal are obtained to calculate the Pearson correlation value.

We compare representational spaces between the first and the last epoch for each layer assuming that it will show the most changes in the model’s inner state.

3.4 Probing

The probing is used to understand whether the model is able to learn the difference between sounds of different languages. We trained a logistic regression classifier for each layer and epoch to predict the language of the sound. The accuracy of the prediction is obtained for each classifier. If the value of accuracy is high then the ASR model encodes information about language of the sound. We compare the accuracies of each layer to understand the learning dynamics.

3.5 Cluster analysis

The process of learning is accompanied by the changes of the representational space in such a way that the data is clustered, meaning that the model generalizes the knowledge. We want to understand how the clusters are formed and how they change during the training.

We track the movements of the clusters. For this purpose, we calculate centroids of each cluster by computing the average vector of each sound for a layer at each epoch. We study the path of the clusters at each layer. We use T-distributed Stochastic Neighbor Embedding (t-SNE) to create 2D projection to visualize the findings.

4. Results

The results are computed on 100 audio recordings from the test dataset: 50 for each language. Overall, there are 4410 sounds. [**number of sounds (table?)**].

4.1 ReStA

The ReStA shows the similarity of 2 representational spaces: after the first epoch of training and after the last one for each layer. The higher the value of ReStA the more similar the spaces are. The results are depicted in Figure 4. We computed the ReStA on Russian and English data separately and on both of them together.

Overall, it is seen that the first layers have changed a little while the last layers differ considerably. Such a behavior is expected as the lower layers are assumed to have general knowledge about the data while the higher layers are responsible for a specific knowledge for a task for which it is fine-tuned. The last five layers have altered the most.

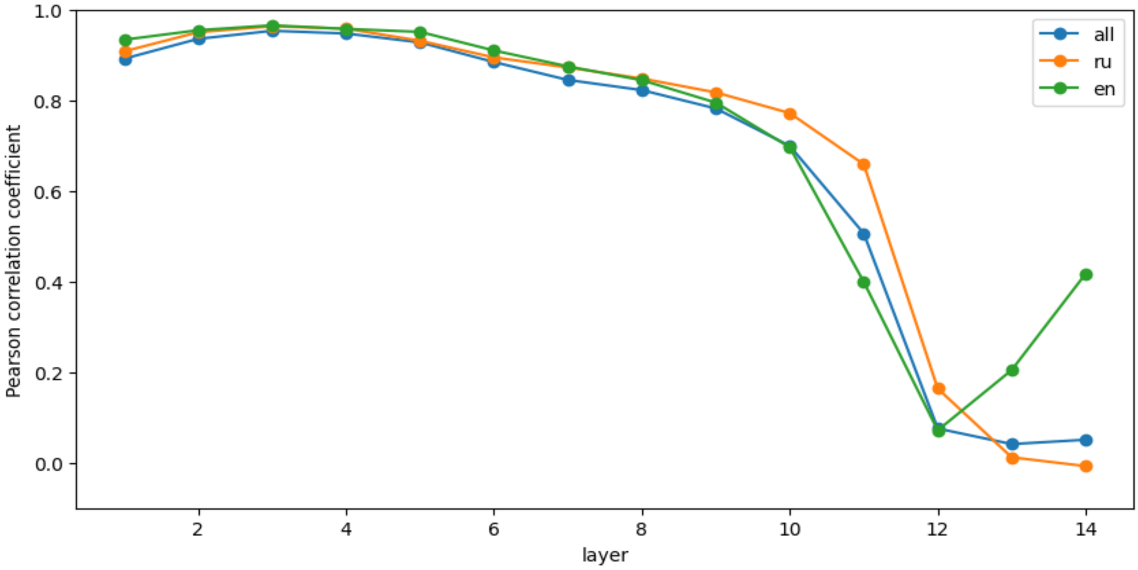


Figure 4. ReStA values between the representations obtained from zero and last epoch for each layer. Green color represents the results on representations of English sounds only, orange color – on Russian sounds only, blue color – on all data. X-axis represents layer number and Y-axis – Person correlation coefficient.

After the 9th layer the Pearson correlation coefficient between English matrices decreases drastically which means that the English sounds representations from 10-12 layers encounter the more changes than the Russian sounds representations from the same layers. Nevertheless, the Russian vectors from the final layers alter the most while the value of correlation on English data increases.

The hypothesis 3 states that the English representations change less than Russian because the wav2vec model was pre-trained on English data and should not relearn it while the Russian audio it hears for the first time. However, the vector alterations are quite similar for Russian and English data. The only deviation from the hypothesis are the results at the 9-12 layers while the results at 13th and 14th layers confirm it.

4.2 Probing

The Figure 5 shows the minimum, maximum and average accuracy among all layers for each epoch. It is seen that at each epoch logistic regression classifier succeeds in predicting the language of a sound. The lowest results are on the zero epoch and on the 10th one. Moreover, almost all minimum accuracies are from the last layer which is a classification layer.

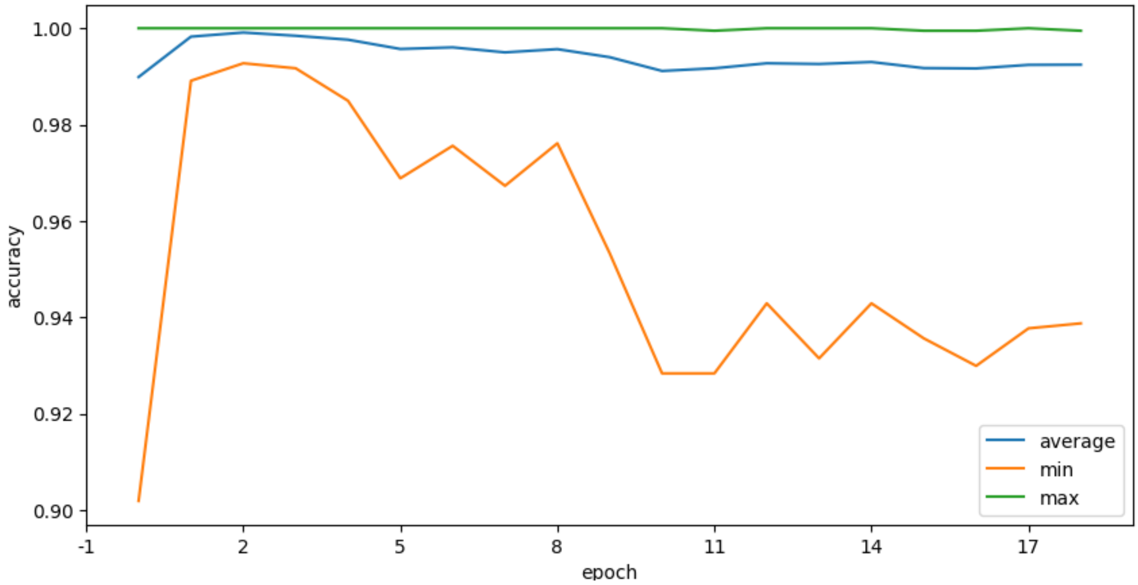


Figure 5. The minimum, maximum and average accuracy among all layers for each epoch. The green line is for maximum accuracies, orange line - for minimum accuracies, blue line – average accuracies. X-axis represents

4.3 Cluster analysis

The paths of cluster centroids are depicted in Appendix A. The graphs show the paths of 5 sounds (t and d at the beginning of a word and between vowels and p at the beginning to show the difference) for each language at each layer. The first notable feature is that sounds on the first three layers move a little which confirms the idea that lower layers almost do not change (§4.1).

The centroids of sounds of the same label tend to be closer to each other, however, the English “d beginning” and “t beginning” centroids move to the opposite direction from the paths of other centroids at 5th, 6th and 7th layer.

In addition, at all layers before the 9th one centroids tend to return to their initial positions which is the state of the model before fine-tuning. After the first epoch of training the centroids become closer to the others of the same language and return back to their previous state after the next epoch. Moreover, these initial positions of English sounds are close to the ones of Russian sounds, meaning that the model is able to capture the similarities between the sounds even before the fine-tuning.

The graph of the 9th layer and subsequent ones depict how the centroids move away from their initial states and from the centroids of other sounds. The same sounds from different languages move in one direction, i.e., pairs of sounds “d”, “t” and “p” at the beginning of the word. Moreover, the path of “d” and “t” centroids are opposite to the path of “p” centroids. Thus, the model encodes the difference between the sounds of distinct nature trying to push the clusters apart. In addition, starting at the 8th layer the centroids of sounds with the same label become closer and create new clusters of “t”, “d” and “p” representations. Generally, these layers are responsible for the alphabetic representations rather than for phonetic ones and this can be seen, for example, in how all “t” labels centroids become closer.

4.4 Phonetic Observations

The phonetic hypothesis 2 states that Russian voiceless sounds are more similar to English voiced sounds at the beginning of a word than to English voiceless sounds as the latter has long lag VOT while the former has similar VOT to Russian voiceless. It should be noted that the English initial voiced consonants have zero VOT while the Russian voiceless stops VOT is a little longer.

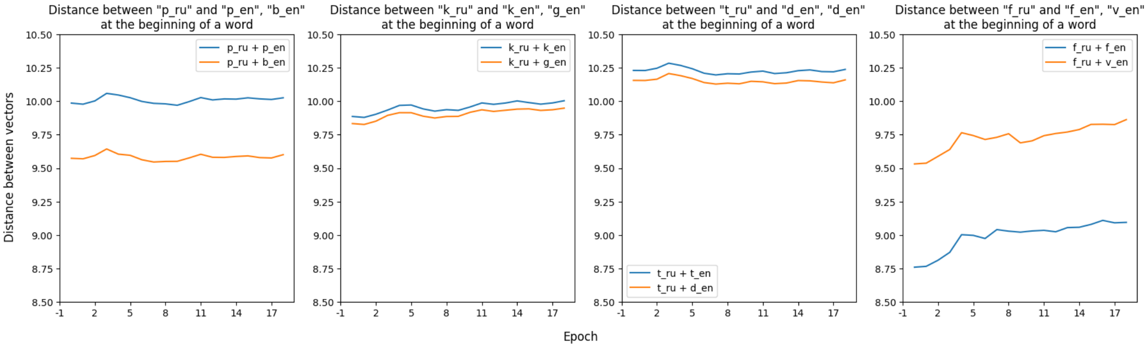


Figure 6. Each graph represents the average distance between the pairs of Russian and English sounds at first layer on each epoch. The orange line is for the pair which is assumed to have less distance than the pair depicted by blue line. X-axis represents the epoch number, Y-axis – distance.

Figure 6 shows the distances between different pairs of labels at first layer on each epoch. The lower the value the closer the vectors to each other. It is noticeable that voiceless Russian stops are closer to voiced consonants in English while it is not true for fricatives “f” and “v”. However, the difference is small for “p”/ “b” pair and even smaller for other pairs of sounds. Thus, the model perceives this peculiarity in the representations of the pairs of sounds “t”/ “d” and “k”/ “g” worse than of the pair “p”/ “b”. Other layers show opposite results.

Figure 7 relates to phonetic hypothesis 1 and the distances between different pairs of labels at first layer on each epoch as the Figure 6. The voiced stops in Russian are more similar to initial voiced stops in English than to the stops between vowels whereas there is an opposite situation for the fricatives. However, similarly to the previous results, the distance difference is relatively small for each group of sounds. Other layers show the same results.

Albeit the main task of the model is to predict the text transcription of an audio, it may learn something about the phonetics as it perceives an audio recording of speech. However, these observations must be checked on the phonetic transcription task to become more accurate.

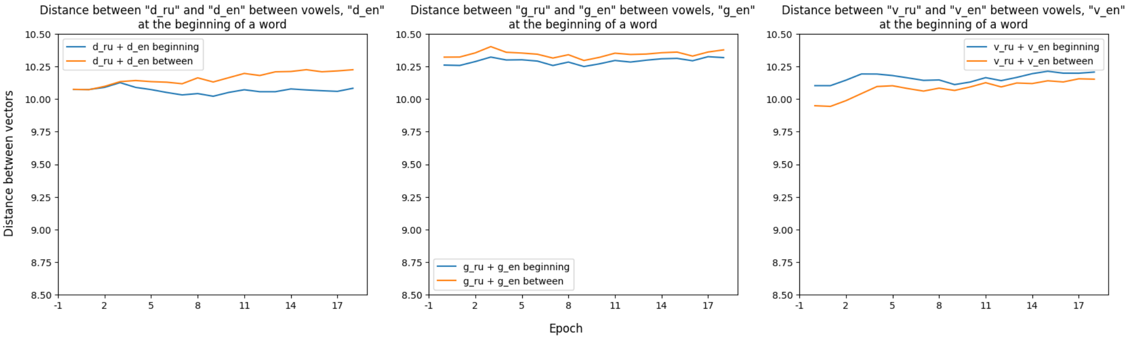


Figure 7. Each graph represents the average distance between the pairs of Russian and English sounds at first layer on each epoch. The orange line is for the pair which is assumed to have less distance than the pair depicted by blue line. X-axis represents the epoch number, Y-axis – distance.

5. Discussion

We want to propose some possible explanations to the obtained results.

In §4.2 it is shown that the model before fine-tuning succeeds in predicting the language of a sound although it has never heard it before. This may be due to the fact that the model at first distinguishes the sounds it already knows from the new ones as non-English sounds. It can be checked by adding audio of another language and conduct the probing test on three languages.

However, there may be other factors which affect the behavior of a classifier and lead to perfect results. For instance, as the recordings are in the different datasets, the model may learn the distinction in the quality of audios.

In §4.3 we show that the centroids tend to get closer after the first epoch of training and then return to the initial state. The assumption is that the model "hears" a new language before fine-tuning and during the first epoch tries to learn the difference. Thus, the model changes the configuration of representational space to make the sounds of one language more similar. Then, it tries to learn relations between languages and return the vectors into the previous state. The higher the layer, the more the final position of the centroid differs from the initial one.

Another observation should be mentioned: the paths of "d_en" and "t_en" in the position between sounds are opposite to the paths of the other sounds of the same label. The possible explanation is that there are a little number of representations for this sounds obtained from the audio thus the centroids may be calculated inaccurately.

6. Conclusion

In this work we conducted several experiments in order to define the process of fine-tuning English Wav2Vec-2.0 on Russian language. In addition, we studied the abilities of the model to perceive phonetic properties of plosive consonants in both languages.

Therefore, the several important features of the model were identified. Firstly, the model might be able to encode the information about the language of a heard sound, however, this is not proved as additional tests should be carried out, for example, probing with more than two languages. Thus, the hypothesis 1 is rejected.

The second hypothesis is also rejected. The model is able to capture the similarities between the sounds of English and Russian language even before the fine-tuning as shown in §4.3.

The third hypothesis is partially confirmed. Although the ReStA show that the English and Russian sounds representations change similarly, at 13th and 14th layers vectors of English sounds change less than the ones of Russian sounds.

The fourth hypothesis is confirmed. The results of ReStA prove that the first nine layers change less than the last ones. In addition, this can be seen on cluster movements graph where at these layers the clusters final position is close to the initial one.

The fifth hypothesis is confirmed on layers 10-14. It is seen that the final positions of clusters centroids are far from the initial positions. Moreover, the vectors of different sounds move away from each other creating more similar space. However, the first 9 layers do not show the same behavior.

As for the phonetic hypotheses, the hypothesis 1 is rejected as described pairs of sounds are farther than expected. The phonetic hypothesis 2 is confirmed on the first layer of the model. However, the difference between the distances of described pairs is small.

Nevertheless, the dataset with phonetic transcriptions in Russian is required to recheck the phonetic results as they are obtained on textual data which can affect the model's comprehension of a phonetic system of a language. Another suggestion is to fine-tune the model with different classification layer combining Latin and Cyrillic symbols and compare the results with the results of this study.

As a continuation of this work, a more profound phonetic study can be made. For instance, there is a correlation between a voicing of a consonant at the end of a word and the length of the preceding vowel in English: the vowels are longer before voiced consonants. However, in Russian all voiced consonants at the end of a word become voiceless. Thus, it is interesting to study how the model treats Russian consonants at the final position in a word.

Moreover, other properties of representational space can be studied. For example, the separability of the space: whether the clusters are linearly or non-linearly separable or not. In addition, it is interesting to check which sounds change more than the others, if there is any.

All in all, multilingual ASR models are a necessity in a modern world. They are also believed to be good for the languages with little data. There is an assumption that the multilingual models, having learned from multiple languages, know the diversity of sounds and thus are able to capture the peculiarities of phonetics of a low-resource language after a few epochs of fine-tuning. Even though the results of such training are good, it is not proved whether the model understands the phonetic structure of a language and its peculiar properties. Therefore, the process of fine-tuning should be studied as it can show what knowledge the model has and what it lacks.

Literature

Datasets

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue. Timit Acoustic-Phonetic Continuous Speech Corpus. // Web Download. Philadelphia: Linguistic Data Consortium, 1993.

Jui Shah, Yaman Kumar Singla, Changyou Chen, Rajiv Ratn Shah. Common Voice: A Massively-Multilingual Speech Corpus // arXiv preprint arXiv:2101.00387. 2020.

Vassil Panayotov, Guoguo Chen, Daniel Povey, Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015. P. 5206-5210.

Models

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever. Robust speech recognition via large-scale weak supervision // arXiv preprint arXiv:2212.04356., 2022

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // arXiv preprint arXiv:2006.11477. 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. 2019.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. T. 29, 2021. P. 3451-3460.

Phonetics

Leigh Lisker, Arthur S. Abramson. A cross-language study of voicing in initial stops: Acoustical measurements //Word, T. 20, №. 3, 1964. P. 384-422.

Ringen Catherine, Kulikov Vladimir. Voicing in Russian Stops: Cross-Linguistic Implications // Journal of Slavic Linguistics, 20, 2004. P. 269–286.

Tetyana Dittmers, Christoph Gabriel, Marion Krause1 and Sevda Topal. Positive transfer from the heritage language? The case of VOT in German/Turkish and German/Russian learners of L3 French, Russian, and English //Universität Hamburg. Retrieved from:
http://wa.amu.edu.pl/L3_workshop/Dittmers_Gabriel_Krause_Topal_2017_L3Poznan.pdf, 2017.

Mikhail Zaikovskii, Ettien Koffi. An Acoustic Phonetic Account of VOT in Russian-Accented English // Linguistic Portfolios, T. 8, №. 1, 2019. P. 7.

Probing

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July, 2018. P. 2126–2136.

Gregor Wiedemann, SteffenRemus, AviChawla, and ChrisBiemann. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. // arXiv preprint arXiv:1909.10430. 2019

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations // arXiv preprint arXiv:1905.06316. 2019

Jui Shah, Yaman Kumar Singla, Changyou Chen, Rajiv Ratn Shah. What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure // arXiv preprint arXiv:2101.00387. 2021.

Fine-tuning analysis

Elena Voita, Rico Sennrich, and Ivan Titov. The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives // In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4396–4406, Hong Kong, China, 2019. P. 4396–4406.

Nikolaus Kriegeskorte, Marieke Mur, Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience // Front Syst Neurosci, 2008.

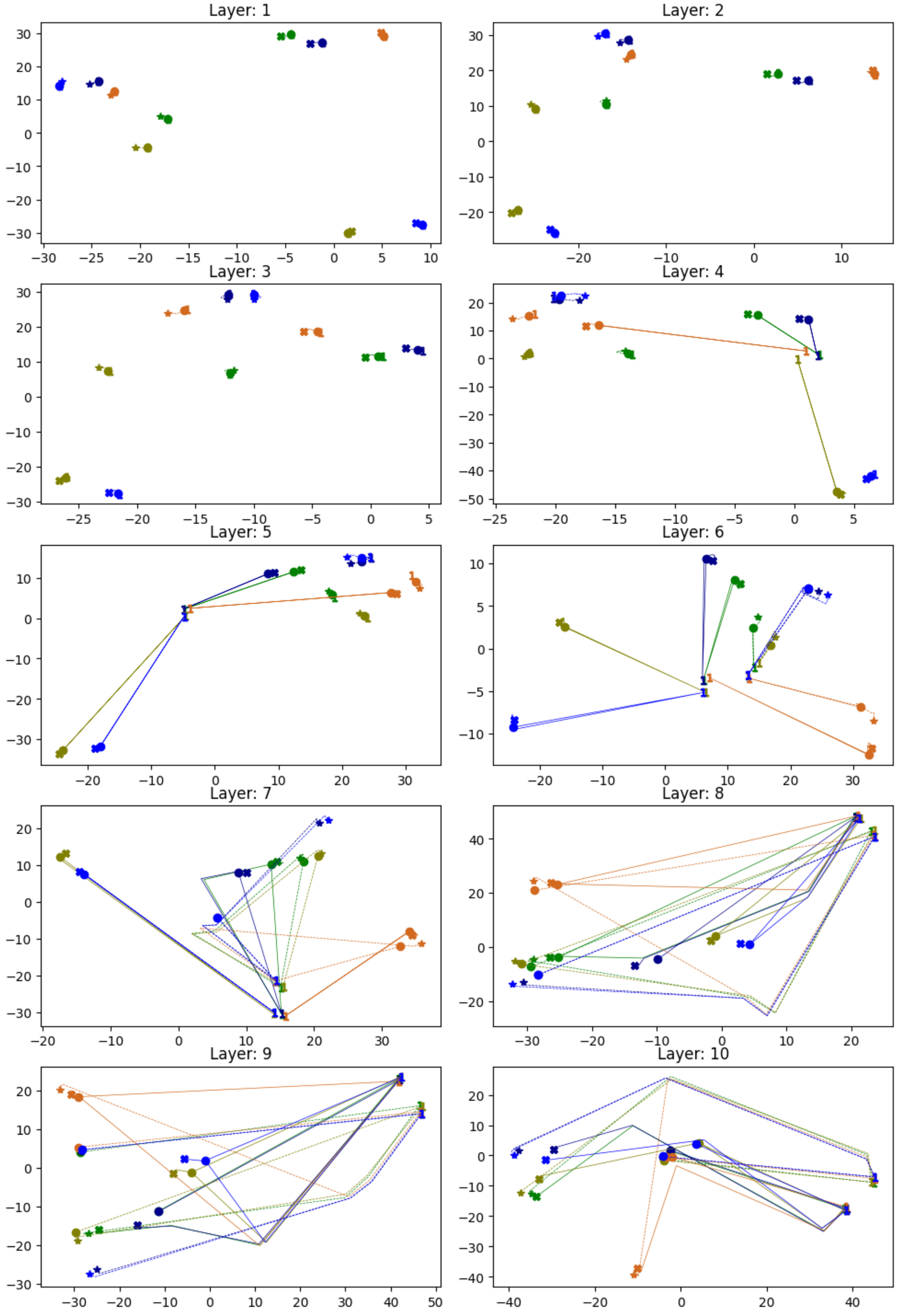
Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains // In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 2019. P.191–203.

Yaru Hao, Li Dong, Furu Wei, Ke Xu. Investigating learning dynamics of BERT fine-tuning // Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020. P. 87-92.

Yichu Zhou, Vivek Srikumar. A closer look at how fine-tuning changes BERT // arXiv preprint arXiv:2106.14282, 2021a.

Yichu Zhou, Vivek Srikumar. DirectProbe: Studying representations without classifiers // arXiv preprint arXiv:2104.05904, 2021b.

Appendix A.



Layer: 1

Layer: 2

Layer: 3

Layer: 4

Layer: 5

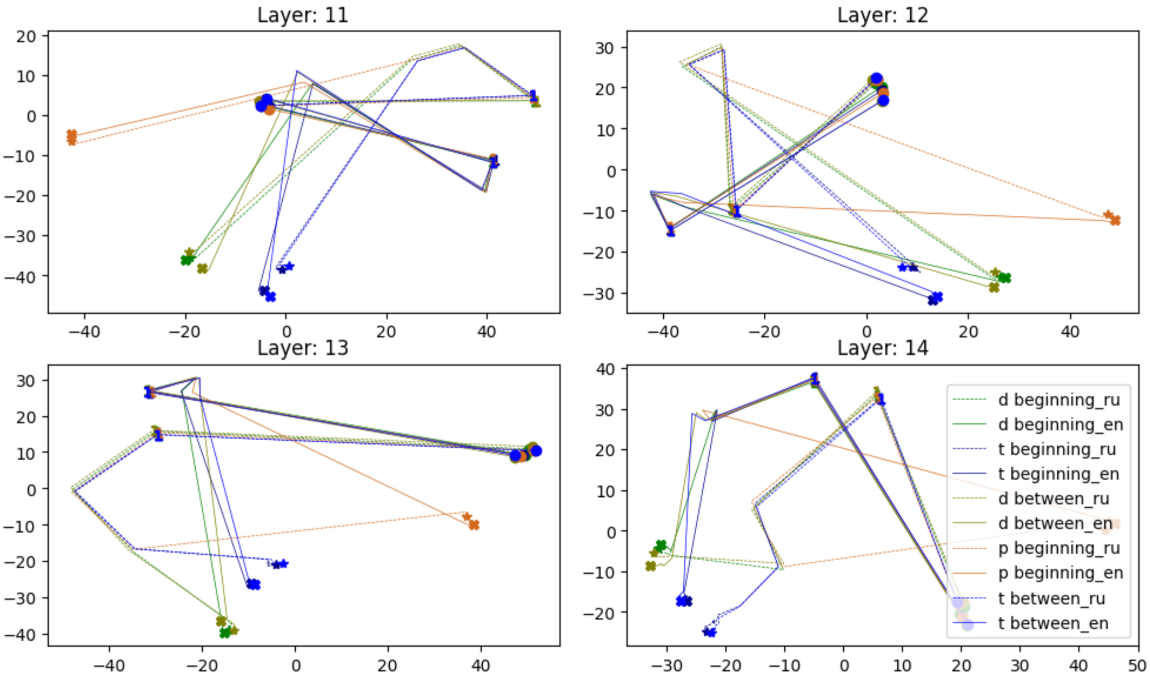
Layer: 6

Layer: 7

Layer: 8

Layer: 9

Layer: 10



Appendix A. The paths of the centroids of [d], [t], [p] sounds with their position in a word at each layer during the fine-tuning. Dotted line is for Russian sounds, solid line is for English ones. The circle marks the centroid position before the training, the X-mark is for the final position of English sounds centroids, the star is for the final position of Russian sounds centroids. Green color is for “d” at the beginning of a word, dark blue – for [t] at the beginning of a word, olive – for [d] between vowels, blue – for [t] between vowels and orange – for [p] at the beginning of a word.

Appendix B

Epoch	**Training Loss**	**Validation Loss**	**Word Error Rate**
1	-	3.237621	1
2	4.5514	3.157347	1
3	3.0936	3.0987	1
4	2.2426	1.191277	0.781061
5	2.2426	0.873133	0.699220
6	1.0575	0.845601	0.667253
7	0.7651	0.783346	0.643854
8	0.6316	0.779965	0.633857
9	0.6316	0.726073	0.605954
10	0.5272	0.769074	0.587169
11	0.4607	0.734299	0.578491
12	0.4316	0.738307	0.562561
13	0.4316	0.729981	0.556520
14	0.3775	0.737511	0.556849
15	0.3505	0.757217	0.548171
16	0.3169	0.743401	0.545864
17	0.3169	0.759330	0.535098
18	0.3112	0.778451	0.536746

Appendix B. Training loss, validation loss and word error rate at each epoch of fine-tuning.

Appendix C.

Sound + position	Russian	English
b beginning	17	19
b between	6	0
d beginning	20	23
d between	17	2
f beginning	1	18
f between	0	1
g beginning	8	12
g between	17	6
k after s	10	8
k beginning	18	3
k between	6	0
p after s	5	4
p beginning	67	12
p between	8	3
t after s	12	8
t beginning	16	56
t between	19	7
v beginning	42	3
v between	14	10
Sum	**385**	**257**

Appendix C. The number of sounds used for testing phonetic hypothesis.