Speech Recognition Interpretation

Ksenia Sherman
HSE University
kvsherman@edu.hse.ru

Oleg Serikov
HSE Univeristy, AIRI, DeepPavlov lab
oserikov@hse.ru

Abstract

Neural Networks show state-of-the-art performance on a number of different NLP tasks. Such behaviour is interesting for AI researchers because it shows that the models may "know" something about the natural language. *Probes* are proposed to assess this. They estimate how the output from the language transformer correlates with different linguistic properties.

Language models are useful not only in domestic tasks but also for scientific research in linguistics, for example, for creating transcription the endangered languages. Thus, such studies may show the limitations of the models which should be overcome in order to enhance the performance.

In this study we probe ASR transformer model Wav2VecPhoneme for three phonetic concepts. We assess the ability of the model to encode different properties of the sound through diagnostic classification. Then we study the embeddings themselves in detail and propose an explanation of the peculiarities in the prediction of the language model. We make our results publicly available.

1 Introduction

Transformer models have drawn a lot of attention since their appearance as they show success in solving many different tasks such as computer vision, image generation, speech reproduction and recognition. These models achieve state-of-the-art results, however, they have poor interpretability. Parametrized networks are difficult to analyse and it is hard to extract any structures and mechanisms which emerge within a model. Thus, the way transformers encode the language remains a mystery.

There are numbers of studies trying to find an explanation of the behavior of models. The most popular method is *probing* which allows to assess the model's knowledge in some areas of interest. The most attention is paid to (Natural Language Processing) NLP models, therefore, a lot of research has been carried out studying the models which produce and analyze written speech but a little work has been made in the field of speech recognition.

In this work we conduct research studying the way the Wav2Vec model captures phonetic properties of a language. We base our study on Diagnostic Classification probing and Behavioral probing trying to describe not only the fact that the model captures (or not) some linguistic information but the way how it creates embeddings and analyses a language.

2 Related Work

Since the very core idea of text embedding models is linguistically motivated (Firth, 1957), attempts to linguistically interpret the vector models have been made. In (Sahlgren, 2006) and in (Rogers et al., 2018) authors study the linguistic concepts emerged in static vector models. In (Handler, 2014), authors wonder if the distributional semantic embeddings follow the semantic ontology structure. Several works concerning the linguistic interpretation of contextual models, translation (Alain and Bengio, 2016), and then language models (Rogers et al., 2020), arose later in the field of NLP.

BERT (Devlin et al., 2019) is arguably the most used pre-trained model in the field of Natural Language Processing. Being originally trained to perform masked language modeling and identify sentences entailment, BERT is believed (Hewitt and Manning, 2019) to capture the morphosyntactic linguistic regularities. It follows the Transformer architecture, consisting of 24 transformer layers in the studied version. Despite grammatical and semantic studies compose the majority of novadays probing studies, phonetic regularities got inspected. Back in 1997, (Rodd, 1997) discover in-

dividual neurons capable of phonetic knowledge such as vowels vs consonants distinction or vowels harmony in character-level language models for Turkish language.

As the interpretability branch of research became popular, *probing* became the prevalent methodology. Introduced in (Adi et al., 2016), and popularized by (Conneau et al., 2018a), the methodology of diagnostic classification serves as the baseline approach for probing studies. Diagnostic classification studies were criticized (Pimentel et al., 2021) for the misleading nature of the accuracy score widely used in studies. Different concepts, including MDL probing technique (Voita and Titov, 2020) and selectivity (Hewitt and Liang, 2019) term were introduced contributing to the reliability of probing methods.

In (Voss et al., 2021), authors reveal the presence of shape- and pattern-specific sub-networks in the popular vision architectures, enriching the body of interpretability works with the visual modality.

In wav2vec-2.0 (Baevski et al., 2020) authors propose the techniques to pre-train models for speech processing. Similar to the masked language modeling objective, they train the model to distinguish the true latent audio representations from the distractor ones.

In (Baevski et al., 2021), authors study how informative are the layers of the model for the task of phonemes classification. They find the middle-to-high layers to be the most informative (yet the quality drops on the very final layers).

3 Methods

We aim to identify whether the transformer gains some information about the internal structure of the sound it hears through the process of learning. Besides, we try to reveal the distribution of this "knowledge" across the layers of the transformer. In addition, we seek for the detailed description of the way the model understands phonetic features of the language.

3.1 Diagnostic Probing

We use probing classifier to define the relationship between embeddings from the output of the model and categories of some phonetic phenomenon (Conneau et al., 2018b; Alain and Bengio, 2017). The layer activations are fed to the logistic regression classifier to predict the language property. The performance of the classifier is measured with accuracy score on each layer of the model: the higher score is, the better some phonetic characteristic is encoded in the embeddings.

However, the behaviour of the classifier is not that transparent. Namely, the high accuracy score may be obtained due to specific training dataset or the ability of the model to adjust to the data. Hence, the use of accuracy only in diagnostic probing has been widely criticized. Control task is proposed to avoid ambiguity in interpretation of the results of probing classifier (Hewitt and Liang, 2019). Such tasks are designed to determine whether the model results are reflected by encoded linguistic properties in the output representations or they are obtained independently. Thus, the desired probe should be *selective*, in other words, should have high accuracy on probing task and low accuracy on control task.

In our work we assess the selectivity of the language model using two control tasks. The first one is *random initialization* (RI): all vectors get random label from the list of labels for each language property. The second one is *grouped random initialization* (GRI): all representations of random sounds change their label to the opposite one. For instance, every embedding of the sound [p] gets a new label 'vowel' instead of 'consonant', while for the first control task some embeddings of the same sound may get 'vowel' label and others — 'consonant'.

3.2 Behavioral Probing

We study the language model's outputs thoroughly to discover the way it learns information from embeddings. This approach helps to control the output because it requires curated datasets, thus, we decrease the field of interest and develop interpretability (Linzen et al., 2016; Goldberg, 2019).

This approach also allows to learn the way the model creates representations and whether there is any specific knowledge within created embeddings.

Both embedding analysis and error analysis, which we describe in the following section, are related to this field of probing.

3.3 Data

This research is based on the data from Common Voice dataset (Ardila et al., 2020) which is located in Hugging-Face repository. For examination we

have chosen russian language as it has interesting sound properties such as sonority and softness. Overall, we have obtained sound vectors from 800 audios and got more than 17000 embeddings from each layer.

There is no markup in the data, thus, we use Wav2Vec labels. It has high quality in prediction, therefore, we make an assumption that this marking is good enough for our study and vectors represent the true sounds.

4 Experiments

The research is divided into three sections: main probing task, embedding analysis and error analysis. It is conducted on the Wav2Vec2Phoneme model (Xu et al., 2021).

4.1 Model

For our experiments we use Wav2Vec2Phoneme model from Hugging-Face repository. It is based on pretrained XLSR-53 model and trained to produce the transcription of an audio.

Embeddings from the output of the model contain 392 numbers each. The position of each number (index) is related to some sound from the vocabulary of the model. For example, the 24^{th} index in the vector is related to the sound [j] and 10^{th} index — to the sound [i]. The higher the number in the vector, the more confident the model is in its prediction. We will refer to a number at some index in a vector as a number of some sound or as a degree of confidence.

4.2 Diagnostic Probing Task

We choose three probing tasks to evaluate the ability of the model to understand phonetics of the language:

- Vowel—Consonant sounds task to predict whether the sound is vowel or consonant.
- Voiced—Voiceless sounds task to determine whether the vocal cords are used to produce a sound.
- Soft—Hard sounds task to classify sounds in terms of tongue location: near soft palate or not.

All representations are given the suitable label and then they are fed to Logistic Regression classifier. This test repeats on all layers and with all control tasks. The first task is the basic one and believed to be the most 'easy' for the model to predict. Whereas the second and the third may be challenging because such a division into two groups is not ubiquitous because some languages do not have these sound properties, thus, the model may not have enough data for training to predict them.

4.3 Embedding Analysis

The degrees of confidence in an embedding allow us to estimate the way the model understands which sounds are similar to the predicted one. Therefore, we try to determine whether the transformer has some knowledge about the sounds with similar realizations, namely, which sounds the model thinks to be the most similar to the one predicted. Thus, we create an average vector for each sound and study the numbers within all such vectors.

4.4 Error Analysis

We explore the errors that the model makes to determine whether there is a pattern in model's mistakes.

In our work we examine the way the language model predicts soft consonants in the last stressed syllable of different words. This location is chosen to avoid ambiguity in the results: the stressed vowels are pronounced clearly, therefore, the possibility of bad prediction based on weak pronunciation is decreased.

Moreover, we assume that though the model has errors in predicting soft consonants, it still captures the information about softness of the sound in representation.

5 Results

5.1 Diagnostic probing task

Overall, figure 1 shows that RI control task has poor results in predicting classes in each of the tests whereas GRI results are quite similar to the accuracy score on correct data. However, the results of probing tasks differ. The following subsections will describe only the GRI control task and correct data accuracies because they are considered to be more interesting for the research whereas RI control task shows high selectivity in each task.

5.1.1 Vowel—Consonant task

The selectivity of Wav2Vec decreases from the first to last layer even though the accuracy on the

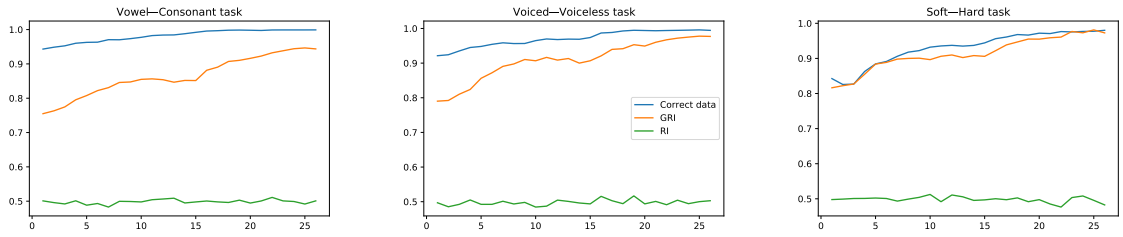

Figure 1: The accuracies of LogReg model's predictions on three probing tasks. The blue line represents the result on correct data, the orange one represents control task on group random initialization and the green one - random initialization. X-axis represents layer number and Y-axis - the accuracy score

correct data increases and reaches a high value of 0.999. The difference between GRI results and correct data results is higher on the initial layers of the network: 18.8% on the first layer. This means that the ASR model better captures the vowel—consonant knowledge in the initial layers of the network.

In addition, we examine coefficients the LogReg model uses for prediction on the last layer: the higher the modulo coefficient, the more important the feature is for prediction. The highest positive score has [i], [oʊ] sounds and the lowest negative number has [j], [m], [l] sounds (Appendix A).

5.1.2 Voiced—Voiceless task

The selectivity in this task differs from the previous one: it fluctuates between 13% and 1,7% that is less than the outcome in vowel—consonant task. Nevertheless, the tendency is the same - the first layers have higher selectivity than the last ones. It leads to the conclusion that the first layers better capture the sonority of sounds. However, this property is more difficult to capture than that of the previous section.

The coefficient analysis shows that the most important for the LogReg prediction are [v], [t] and [sʲ] sounds with positive coefficients and [b], [r], [i] sounds with negative coefficients (Appendix A).

5.1.3 Softness—Hardness task

The results of this probing task distinct from both previous tasks in a way that the selectivity nearly equals zero or is even negative on some layers. The low accuracy on the correct data is due to the labeling which is based on the Wav2Vec predictions. The model makes mistakes and some sounds which are labeled as hard sounds are in fact soft. Nevertheless, the LogReg model is able to detect the differences between the embeddings of soft and hard sounds and predicts right labels. This leads to worse results on the correct data comparing to other tasks. Thus, the language model is not accurate enough in its predictions but still keeps some information about the softness of a sound within a vector. More detailed research is in the section 5.3.

The [tʃ], [j] sounds have the highest positive LogReg coefficients and the lowest negative coefficients are received by [l], [z] (Appendix A).

5.2 Embedding Analysis

Examining average vectors of the sounds we obtained different results. First, the realizations of the sounds which get the highest number in the embedding are similar to the realization of the predicted sound. Hereby, vowels have the highest degree of confidence in the vectors of vowels and the consonants - in vectors of consonant sounds. Moreover, the articular properties of the predicted sound are mostly similar to the articular properties of the ones with high degree of confidence. For example, in the embedding of [m] sound the sound [n] has high score as being a nasal sound. Another example is that soft sounds get high score in the embeddings of other soft sounds. Likewise, consonants which differ only in the use of vocal cords, for example, [t] and [d], have high degree of confidence in the embeddings of their counterparts.

A heat map which represents the way the numbers are distributed within a vector can be found in the Appendix B.

5.3 Error Analysis

The results of softness-hardness task show that the Wav2Vec model has difficulties in predicting soft consonants. Even though we cannot suggest that there is no information about this property captured within the embeddings, we see the huge difference between the results of softness-hardness

probing task and the others.

The analysis of soft sounds' vectors from the last layers demonstrates that their embeddings get high number at the index of [j] sound and other soft sounds. Therefore, we calculate the distribution of the [j] sound degrees of confindence and take these sounds which are outliers. Thus, this examination illustrates that the resulting representations are more like the representations of soft sounds because they have high numbers at the indexes of sounds [j], [ʃ], [ɲ] and others which are claimed to be soft.

In addition, the predicted sound has the highest number in the embedding and this number is usually more than the rest by four. However, the prediction of soft sounds is not that confident and the highest number differs from others by no more than two. The embeddings of the outliers also show this tendency.

The closer look on the predictions allows us to make some new assumptions about the way the model makes mistakes. To begin with, the model never predicts [pʲ], [bʲ], [kʲ], [gʲ] sounds. Some sounds are predicted as soft better than others, for instance, palatalized [nʲ], [tʲ] and [dʲ] are predicted more often than their hard pairs. Sometimes model predicts hard sounds and [j] as separate sounds, for example, [rju]. However, there is no visible pattern in the way the model makes mistakes because it predicts different sounds in the same word or in the same phonetic environment, for instance, /blʌgʌderu/ and /blʌgʌderʲu/.

6 Conclusion

In this work we describe three studies on Wav2Vec model. Firstly, we try to examine the readability of information in a vector with diagnostic probing method. As a result, we conclude that the ASR model better captures vowel—consonant property and the information about the sonority of a sound on first layers. Softness-hardness task reveals that the Wav2Vec model has difficulties in prediction this property which affects the accuracy score on the probing task.

Secondly, we investigate the embeddings to understand the way the transformer creates them. We determine that the model gives high degree of confidence to sounds which are more similar to the predicted one. Namely, the model have an understanding which sounds have similar realization.

Finally, we study at the mistakes the Wav2Vec model makes when it predicts consonants. It reveals that even though the model predicts hard sounds, the embedding has the information that the sound is soft. It is illustrated by the fact that soft sounds in the vector get the high degree of confidence. Moreover, the model better predicts palatalized [nʲ], [tʲ] and [dʲ] sounds whereas [pʲ], [bʲ], [kʲ], [gʲ] sounds do not appear in the prediction.

One of the possible explanation to such a behavior is that the model did not have enough data to learn the softness of sounds as small set of languages has this property.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *CoRR*, abs/2105.11084.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

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

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

Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv e-prints*, page arXiv:1901.05287.

Abram Handler. 2014. An empirical study of semantic similarity in wordnet and word2vec.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *ArXiv*, abs/1909.03368.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tiago Pimentel, Brian Roark, Søren Wichmann, Ryan Cotterell, and Damián Blasi. 2021. Finding concept-specific biases in form–meaning associations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4416–4425, Online. Association for Computational Linguistics.

Jennifer Rodd. 1997. Recurrent neural-network learning of phonological regularities in Turkish. In *CoNLL97: Computational Natural Language Learning*.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *Linguistics*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *CoRR*, abs/2003.12298.

Chelsea Voss, Gabriel Goh, Nick Cammarata, Michael Petrov, Ludwig Schubert, and Chris Olah. 2021. Branch specialization. *Distill*. Https://distill.pub/2020/circuits/branch-specialization.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. *arXiv e-prints*, page arXiv:2109.11680.

A Appendix

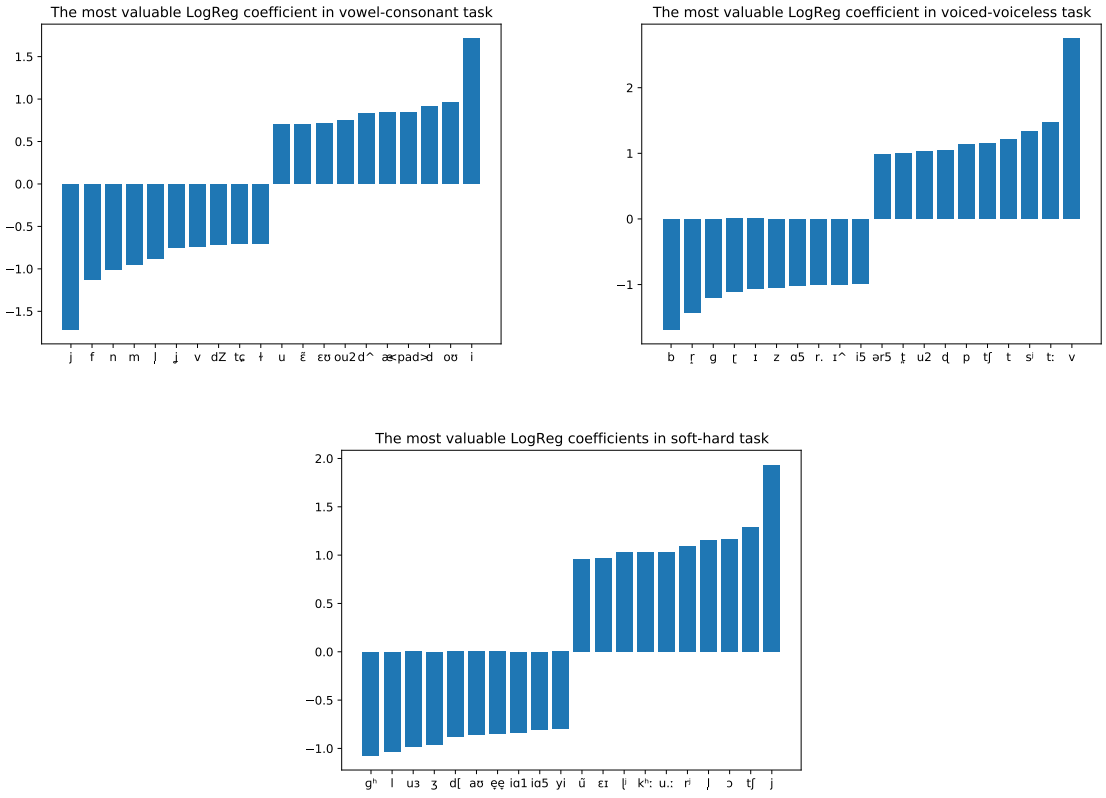

Figure 2: The bar charts illustrate the most important for prediction coefficients which the Logistic Regression model has after training. X-axis represents the sounds and the Y-axis represents the coefficient values.

B Appendix

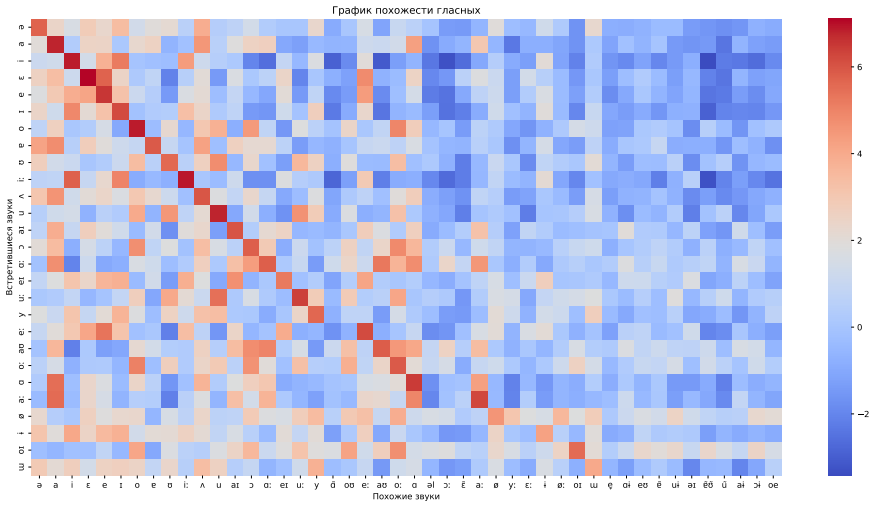

Figure 3: The heatmap shows the degree of vowel similarities based on the embeddings' degrees of confidence. The Y-axis represents the sounds which are in the vocabulary of the model. The X-axis represents all the sounds which are the most similar to the ones in the vocabulary. The redder the colored square, the more similar the two sounds are (the higher degree of confidence the sound from y-axis has at the index of a sound from x-axis).

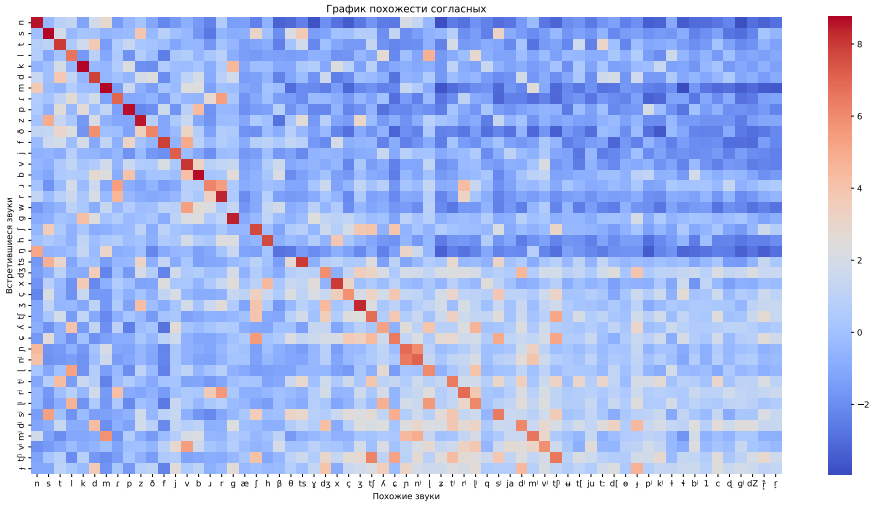

Figure 4: The heatmap shows the degree of consonant similarities based on the embeddings' degrees of confidence. The Y-axis represents the sounds which are in the vocabulary of the model. The X-axis represents all the sounds which are the most similar to the ones in the vocabulary. The redder the colored square, the more similar the two sounds are (the higher degree of confidence the sound from y-axis has at the index of a sound from x-axis).