УДК 004.852
DOI: 10.26102/2310-6018/2024.44.1.002

Dysarthria speech recognition by phonemes using hidden Markov models

B.A. Bredikhin[1,2]✉, M.H. Antor[1], N.A. Khlebnikov[1], A.V. Melnikov[1], M.V. Bachurin[1]

[1]Ural Federal University, Yekaterinburg, the Russian Federation
[2]CyberLympha, Skolkovo, the Russian Federation

Abstract. The relevance of the paper is due to the difficulties of oral interaction between people with speech disorders and normotypic interlocutors as well as the low quality of abnormal speech recognition by standard speech recognition systems and the inability to create a system capable of processing any speech disorders. In this regard, this article is aimed at developing a method for automatic recognition of dysarthric speech using a pre-trained neural network for recognizing phonemes and hidden Markov models for converting phonemes into text and subsequent correction of recognition results using a search in the space of acceptable words of the nearest Levenshtein word and a dynamic algorithm for splitting the output of the model into separate words. The main advantage of using hidden Markov models in comparison with neural networks is the small size of the training data set collected individually for each user, as well as the ease of training the model further in case of progressive speech disorders. The data set for model training is described, and recommendations for collecting and marking data for model training are given. The effectiveness of the proposed method is tested on an individual data set recorded by a person with dysarthria; the recognition quality is compared with neural network models trained on the data set used. The materials of the article are of practical value for creating an augmented communication system for people with speech disorders.

Keywords: hidden Markov models, dysarthria, automatic speech recognition, phonemes recognition, phoneme correction.

For citation: Bredikhin B.A., Antor M.H., Khlebnikov N.A., Melnikov A.V., Bachurin M.V. Dysarthria speech recognition by phonemes using hidden Markov models. *Modeling, Optimization and Information Technology*. 2024;12(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1471 DOI: 10.26102/2310-6018/2024.44.1.002

Распознавание дизартричной речи по фонемам с использованием скрытых марковских моделей

Б.А. Бредихин[1,2]✉, М.Х. Антор[1], Н.А. Хлебников[1], А.В. Мельников[1], М.В. Бачурин[1]

[1]Уральский федеральный университет, Екатеринбург, Российская Федерация
[2]ООО «Сайберлимфа», Сколково, Российская Федерация

Резюме. Актуальность работы обусловлена сложностями устного взаимодействия людей с нарушениями речи с нормотипичными собеседниками, а также низким качеством распознавания аномальной речи стандартными системами распознавания речи и невозможностью создания системы, способной обработать любые нарушения речи. В связи с этим данная статья направлена на разработку метода автоматического распознавания дизартричной речи с применением предобученной нейронной сети для распознавания фонем и скрытых марковских моделей для преобразования фонем в текст и последующей коррекции результатов распознавания с помощью поиска в пространстве допустимых слов ближайшего по расстоянию Левенштейна слова и динамического алгоритма разбиения выхода модели на отдельные слова. Основное преимущество использования скрытых марковских моделей по сравнению с нейронными сетями заключается в малом размере обучающего набора данных, собираемого индивидуально для каждого пользователя, а также в простоте дообучения модели в случае прогрессирующих

нарушений речи. Описывается набор данных для обучения модели, и даются рекомендации по сбору и разметке данных для обучения модели. Эффективность предложенного метода проверяется на индивидуальном наборе данных, записанных человеком с дизартрией; качество распознавания сравнивается с нейросетевыми моделями, обученными на используемом наборе данных. Материалы статьи представляют практическую ценность для создания средства дополненной коммуникации для людей с нарушениями речи.

Ключевые слова: скрытые марковские модели, дизартрия, автоматическое распознавание речи, распознавание фонем, коррекция фонем.

Для цитирования: Бредихин Б.А., Антор М.Х., Хлебников Н.А., Мельников А.В., Бачурин М.В. Распознавание дизартричной речи по фонемам с использованием скрытых марковских моделей. *Моделирование, оптимизация и информационные технологии.* 2024;12(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1471 DOI: 10.26102/2310-6018/2024.44.1.002 (на англ.)

Introduction

It is difficult for people with speech disorders to communicate with normotypic interlocutors. Common voice assistants do not appear to be helpful due to the poor quality of speech recognition. The variety of speech disorders, especially hyperkinetic dysarthria which does not have certain patterns, does not allow creating a single automatic speech recognition system suitable for a large number of people with speech disorders [1]. That is, it is necessary to collect an individual data set for each user to train the recognition model. However, collecting a large data set required to use neural networks is often difficult because of the slow pace of speech and/or rapid fatigue of the target user. Also, if the individual features of speech change over time or have a probabilistic nature, it will be necessary to collect a new data set comparable in volume to the original.

There are publications on the recognition of abnormal speech which address mainly native English speakers. According to [2], either artificial neural networks or hidden Markov models are generally used for the analysis of dysarthric speech.

Xiong and Barker used articulatory based data having WER about 0,49 [3]. Nevertheless, this approach uses MRI data instead of voice; therefore, it cannot be used in a casual environment. These authors also employ GMM-HMM approach for speech recognition [4] which is similar to the proposed method and requires a small amount of training data. At the same time, there is no human-feedback stage that is useful for model fine-tuning, and the method shows WER about 0,69.

Two papers were taken as the basis of this article, one of which provides the solution to the problem of recognizing predefined phrases, and the second examines automatic diagnostics of the degree of speech disorders.

Hidden Markov models is used by Hawley [5] to translate MFCC spectrograms into a hidden state, which is classified by a small (up to 50 elements) dictionary as one of the predefined phrases. To train the system, it took from 600 to 2000 audio recordings for each user.

The translation of speech into phonemes using the XLS-R neural network and comparison of the result with the reference phrase using the loss function is used in [6]. There is no direct speech recognition in this paper, but it is possible to recognize text based on the features extracted by XLS-R.

The paper proposes a hybrid model combining an artificial neural network and a hidden Markov model. Answers to the following questions are given:

1. Is it possible to use a pre-trained multilingual phoneme recognition model?
2. Is it possible for a person to give meaningful feedback for directed training of a speech recognition model?

3. What is the minimum data set size required for high quality speech recognition?

4. Is it possible, having trained the model on a small set of phrases, to recognize words from a large dictionary.

Materials and methods

For the purposes of speech recognition model training and evaluation, a data set of 2000 Russian phrases spoken by a man with hyperkinetic dysarthria were recorded [7].

Hyperkinetic dysarthria is characterized by random voice flow stops, including stops for breathing, variable rhythm and tempo [1]. Each phoneme can be realized by a wider set of sounds, which is different from the norms of the Russian standard language. It is also possible to implement separate combinations of two phonemes by combining more sounds. These features make speech recognition difficult by standard methods.

To complete the data set, phrases of various subjects are recorded: typical personal conversations, pleas, study materials. The phrases were recorded sequentially from September 2022 to May 2023. The length of phrases varies from 1 to 718 words, the total duration of audio recordings is 2 hours 46 minutes 15 seconds, the minimum duration of one recording is 1 second, the average is 5 seconds, the maximum is 868 seconds.

For phoneme recognition, a pre-trained XLS-R neural network by Xu et al [8] is chosen. This neural network is trained on audio recordings and their phonetic transcriptions of the Common Voice and Babel data sets in 45 languages of various language families. In total, the model is able to recognize 392 phonemes. Figure 1 shows the results of audio recording recognition from the data set described in the "Data collection" item.

5	это не имеет значения	e l t ʌ n i m e t u z y z n ʌ ʧ e ɲ i g ʌ
6	буквально	b o k ʌ v a: l i t s n ɑ
7	кроме	k ʌ t r o m e
8	отпечаток пальца	a t f i ʧ ɛ t a k p a l t s ə
9	зверёк	ð v i k r j ɔ k
10	животные	ʒ ø v ɔ t n a: j e

Figure 1 – Example of phonemes recognized by XLS-R from the data set
Рисунок 1 – Примеры фонем из набора данных, распознанных с помощь XLS-R

As seen above, the recognized phonetic transcriptions roughly correspond to the correct pronunciation; however, there are anomalies, for example, extra sounds ("ʌ t" in the word “кроме”/”krome”/”except”), substitutions of sounds ("o" instead of "u" in the word “буквально”/”bukval'no”/“literally”), omissions of sounds (in the word “имеет”/”imeet”/"it has" two sounds are missing: "j e"). All three types of anomalies can be calculated by means of the Levenshtein distance [9], which is used in next section.

Phoneme recognition can be described as (1)

$$\hat{x} = XLS_R(x), \quad (1)$$

where x is a sound wave; XLS_R is an XLS-R neural network; $\hat{x}$ is a probability matrix of recognized phonemes.

We can formalize phonemes-to-text translation as a hidden state revealing task.

Let us assume that P is a set of phonemes; L is a set of Russian letters; x is a sound wave; $\hat{x}_k$ is a recognized phonemes matrix with shape of $T_1 \times |P|$; y_k is the true text matrix with the shape of $T_2 \times |L|$; $k = \overline{1, K}$ is a sentence index in the data set; $\hat{x}_{i,j}$ is a probability that i-th phoneme is j; $y_{i,j}$ is a probability that i-th letter is j (0 or 1).

Let us assume that $HMM = \langle L, P \times P, \pi \rangle$ is a hidden Markov model; π is a transition matrix $(P \times P) \times L$ at start filled with 0's. L is a state set, all the states are final.

The following algorithm has been used to train the model:

I. $t = 0$.

II. for each $i = \overline{1, K}$:

1. For each $j = \overline{1, min^{-1}}$:

a) $\pi_{x_{i,j}, x_{i,j+1}, L} \coloneqq \pi_{x_{i,j}, x_{i,j+1}, L} \cdot \frac{t}{t+1}$;

b) $\pi_{x_{i,j}, x_{i,j+1}, y_{i,j}} \coloneqq \frac{\pi_{x_{i,j}, x_{i,j+1}, y_{i,j}} \cdot t + 1}{t+1}$;

c) $t \coloneqq t + 1$.

We have the transition matrix corresponding to the phonetic distortions of the person for whom the model is being trained. This model is able to correct one of three phonetic anomalies – the replacement of sounds. Two other errors, extra sounds and missing sounds are corrected at the text recovery stage.

To recover raw text by phonemes sequence P, a T matrix of size $|L| \times T_2$ is constructed as shown in equation (2):

$$T_{x,t} = \pi(P_t, P_{t+1}, x). \quad (2)$$

To recover text, the method requires a list of acceptable words or phrases W which is stored as a text file or CSV-file.

1. Let us assume that $T_{x,t}$ is a letter probability matrix (2); $\hat{r} = ''$ is a resulting text; $st = 0$ is a current word start index; $mw =, md = 0$ are the next most probable word and its similarity index; $d(s_1, s_2) = 1 - \frac{LD(s_1, s_2)}{|s_1||s_2|}$ is character error rate; LD is a Levenshtein distance.

2. $mw =, md = 0$.

3. $i = |r|$.

4. For each word $w \in W$:

a) $d_1 = d(r[st: i], w)$;

b) if $d > md$, then $md = d, mw = w, st_1 = i$;

c) $st = st_1, \hat{r} = \hat{r} \oplus mw$, $\oplus$ is string concatenation with space.

5. $i \coloneqq i - 1$.

6. If $st \geq |r|$, go to step 7.

7. If $i > st$, go to step 4.

8. $\hat{r}$ is the final text.

Results

The method has been tested on the data set [7] split into a train set of 1 335 phrases and a test set of 445 phrases. For model evaluation purposes, the list of all data set phrases were used as a list of acceptable phrases W.

To score models, the character error rate (CER) metric (lower is better) has been used, which is a Levenshtein distance [9] divided by a string length.

The proposed method has CER of 0,39 on train set and 0,36 on a test set.

Original wav2vec2 model [10] has CER of 0,97, and a custom combination of CNN and LSTM has CER of 0,37, which is comparable to the proposed method, but requires more computational resources and does not support online learning.

Our method supports online learning with the algorithm described in the section “Materials and methods”. To implement the method, a web interface with speech recorder, phonemes field, result field and human feedback field was built. Human feedback contains only those letters which are presented by the recognized phonemes.

To evaluate online learning, first several lines of original Russian text of “Eugene Onegin” by Alexander Pushkin were used. Word list W in this case was constructed from the words used in selected lines. As shown in Table 1, the proposed method can correctly recognize dysarthric speech after 10 sentences marked-up with feedback corrections. The feedback string must have the same length as the output of the model, that is, it does not contain letters omitted by the model, and the extra recognized letters must be replaced with the "_" label.

This is the log of human feedback learning:

The first sentence “Мой дядя самых честных правил,” has recognized phonemes “*m aɪ n d e d e s a n a x r e θ n ɔ x p p r a v i*” and recognized text “*мтнпспспаннхреснлхррaоа*”; therefore, CER is 0,58. We correct the recognition result with a text ‘мойдядясамыхчесныхпправи”, CER=0,06. After correction and model fine-tuning recognition result is “момдядясаныхрясймхппрыв” with CER=0,375 after one batch.

Further, model training log is presented in Tables 1, 2.

Table 1 – Log of model training
Таблица 1 – Журнал обучения модели

Epoch	Text	Recognized phonemes	Recognized text	CER of recognized text
1	Мойдядясамыхчестныхправил,	*m aɪ n d e d e s a n a x r e θ n ɔ x p p r a v i*	*мтнпспспаннхреснлхррaоа*	0,58
2	когданевшуткузанемог	*k a g d a n e f ʃ o t k o l z e n e m o k*	*каопаняошвдкбязямямв*	0,6
3	онуважатьсебязаставил	*o n o v a ʒ a t a b e z e θ t a l f y*	*умввййытабезастолв*	0,54
4	илучшевыдуматьнемог	*i l o t f e u v e d d o m a t n e m o g*	*илвтвеувяддоматнамо*	0,42

Table 1(extended)
Таблица 1 (продолжение)

Epoch	Text	Recognized phonemes	Recognized text	CER of recognized text
5	егопримердругимнаука	: *j e v o p k r i v m aɪ k r u p n ts u: p k r o g i m p n a: o k a*	*йевопкчивмокчвпнявпкаугимпнаук*	0,48
...	...	...	...	...
10	полуживогозабавлять	*p a n o ʒ o ð o v a z e b a v u l n e t*	*поножозогозебоввлне*	0,47

Table 2 – Log of model training (corrected text)
Таблица 2 – Журнал обучения модели (после коррекции текста)

Epoch	Correction text	CER of correction text	Recognized text after correction	CER of recognized text after correction
1	мойдядясамыхчесныхпп рави	0,06	момдядясаныхрясймхпп рыв	0,375
2	когданевшутку_занемог	0,02	кагдоневшотквлзянемо	0,3
3	онуважатебязаств_ви	0,15	уноважотвбазесталв	0,43
4	илутшеввыддуматнемог	0,13	илотвыввaддоматнемо	0,36
5	йегопрриммекр_____д_ ругим_наука	0,21	ыемурррригмекчапняяпкс угимднаук	0,48
...	...	...	...	...
10	поножововазабав_л_ят	0,33	паножосугозябомулне	0,52

As we can see, after 10 sentences the proposed method gives us a result of CER 0,47, which is better than wav2vec2 model and is comparable to a neural network trained on 2000 sentences for several epochs.

Conclusion

The paper proposes an effective adaptive algorithm for recognizing abnormal speech based on a hybrid neural network model and Markov models. The effectiveness of the algorithm when being trained on a pre-recorded data set and when being trained with human feedback is shown. The suggested algorithm will be used to develop a voice assistant application for people with speech disorders.

Returning to research questions, we can argue that:

1. It is possible to use a pre-trained multilingual phoneme recognition model because a human can understand the model output and 0,61 of phrases the model predicted can be reconstructed automatically into correct phrases.

2. As we can see in Table 1, it is possible for a person to give meaningful feedback for directed training of a speech recognition model with some rules of phrase to model output alignment. The feedback string must have the same length as the output of the model, that is, it does not contain letters omitted by the model, and the extra recognized letters must be replaced with the "_" label.).

3. The algorithm can start give meaningful recognition result after 10 attempts of human feedback for recognized sentences.

4. Recognizing words from a large dictionary with a model trained on small data is likely to be impossible for hyperkinetic dysarthria but can be effective for less variative speech disorders like stuttering.

Further, we plan to collect a larger data set from speakers with different speech disorders and to provide base models for common types of speech disorders. Individual training of these models will require even less labeled data.

REFERENCES

1. Rowe H.P., Gutz S.E., Maffei M.F., Tomanek K., Green J.R. Characterizing dysarthria diversity for automatic speech recognition: a tutorial from the clinical perspective. *Front. Comput. Sci.* 4:770210. DOI: 10.3389/fcomp.2022.770210.
2. Balaji V., Sadashivappa G. Speech disabilities in adults and the suitable speech recognition software tools – a review. In: *2015 International Conference on Computing and Network Communications (CoCoNet), Trivandrum, India, 2015.* p. 559–564. DOI: 10.1109/CoCoNet.2015.7411243.
3. Xiong F., Barker J., Christensen H. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. *Speech Communication; 13th ITG-Symposium, Oldenburg, Germany, 2018.* p. 1–5.
4. Xiong F., Barker J., Christensen H. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.* p. 5836–5840. DOI: 10.1109/ICASSP.2019.8683091.
5. Hawley M.S., Cunningham S.P., Green P.D., Enderby P., Palmer R., Sehgal S., et al. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering.* 2013;21(1):23–31.
6. Yeo E.J., Choi K., Kim S., Chung M. Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning. In: *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.* p. 1–5. DOI: 10.1109/ICASSP49357.2023.10094605.
7. Hashan A.M., Bredikhin B. Russian Voice Dataset. Kaggle. URL: https://www.kaggle.com/dsv/5954738 (accessed on 12.08.2023).
8. Xu Q., Baevski A., Auli M. Simple and effective zero-shot cross-lingual phoneme recognition. arXiv; 2021. URL: http://arxiv.org/abs/2109.11680 (accessed on 18.05.2023).
9. Levenshtein, V., Binary codes capable of correcting deletions, insertions and reversals. *Doklady AN USSR.* 1965;163(4):845–848. (In Russ.).
10. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems.* Curran Associates, Inc.; 2020. p. 12449–12460. DOI: 10.48550/arXiv.2006.11477.

СПИСОК ИСТОЧНИКОВ

1. Rowe H.P., Gutz S.E., Maffei M.F., Tomanek K., Green J.R. Characterizing dysarthria diversity for automatic speech recognition: a tutorial from the clinical perspective. *Front. Comput. Sci.* 4:770210. DOI: 10.3389/fcomp.2022.770210.
2. Balaji V., Sadashivappa G. Speech disabilities in adults and the suitable speech recognition software tools – a review. In: *2015 International Conference on Computing and Network Communications (CoCoNet), Trivandrum, India, 2015.* p. 559–564. DOI: 10.1109/CoCoNet.2015.7411243.
3. Xiong F., Barker J., Christensen H. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. *Speech Communication; 13th ITG-Symposium, Oldenburg, Germany, 2018.* p. 1–5.
4. Xiong F., Barker J., Christensen H. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.* p. 5836–5840. DOI: 10.1109/ICASSP.2019.8683091.
5. Hawley M.S., Cunningham S.P., Green P.D., Enderby P., Palmer R., Sehgal S., et al. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2013;21(1):23–31.
6. Yeo E.J., Choi K., Kim S., Chung M. Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning. In: *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.* p. 1–5. DOI: 10.1109/ICASSP49357.2023.10094605.
7. Hashan A.M., Bredikhin B. Russian Voice Dataset. Kaggle. URL: https://www.kaggle.com/dsv/5954738 (дата обращения: 12.08.2023).
8. Xu Q., Baevski A., Auli M. Simple and effective zero-shot cross-lingual phoneme recognition. arXiv; 2021. URL: http://arxiv.org/abs/2109.11680 (дата обращения: 18.05.2023).
9. Левенштейн ВИ. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Докл. АН СССР*. 1965;163(4):845–848.
10. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems.* Curran Associates, Inc.; 2020. p. 12449–12460. DOI: 10.48550/arXiv.2006.11477.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Бредихин Борис Андреевич, магистрант, Институт фундаментального образования, Уральский федеральный университет, Российская Федерация.
e-mail: Boris.Bredikhin@urfu.me
ORCID: 0009-0005-7370-9947

Boris A. Bredikhin, Master's Student, Institute of Fundamental Education (InFO), Ural Federal University, Yekaterinburg, the Russian Federation.

Антор Махамудул Хасан, аспирант, ассистент, Институт фундаментального образования, Уральский федеральный университет, Российская Федерация.
e-mail: hashan.antor@gmail.com
ORCID: 0000-0001-7926-9245

Antor M. Hashan, Postgraduate Student, Assistant Lecturer, Institute of Fundamental Education (InFO), Ural Federal University, Yekaterinburg, the Russian Federation.

Хлебников Николай Александрович, директор, Институт фундаментального образования, Уральский федеральный университет, Российская Федерация.
e-mail: na.khlebnikov@urfu.ru

Nikolai A. Khlebnikov, Director, Institute of Fundamental Education (InFO), Ural Federal University, Yekaterinburg, the Russian Federation.

Мельников Александр Валерьевич, студент, Институт радиоэлектроники и информационных технологий, Уральский федеральный университет, Российская Федерация.
e-mail: sanek.melnikov@mail.ru

Aleksandr V. Melnikov, Undergraduate Student, Institute of Radioelectronics and Information Technologies, Ural Federal University, Yekaterinburg, the Russian Federation.

Бачурин Матвей Владимирович, студент, Институт радиоэлектроники и информационных технологий, Уральский федеральный университет, Российская Федерация.
e-mail: matvey_1703@mail.ru

Matvey V. Bachurin, Undergraduate Student, Institute of Radioelectronics and Information Technologies, Ural Federal University, Yekaterinburg, the Russian Federation.

Статья поступила в редакцию 02.11.2023; одобрена после рецензирования 04.12.2023; принята к публикации 17.01.2023.

The article was submitted 02.11.2023; approved after reviewing 04.12.2023; accepted for publication 17.01.2023.