Information systems. Computer sciences. Issues of information security
Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.522
https://doi.org/10.32362/2500-316X-2025-13-3-7-20
EDN PVYBDD

RESEARCH ARTICLE

Accent conversion method with real-time voice cloning based on a non-autoregressive neural network model

Vladimir A. Nechaev [@], **Sergey V. Kosyakov**

Ivanovo State Power Engineering University, Ivanovo, 153003 Russia
[@] *Corresponding author, e-mail: nechaev@gapps.ispu.ru*

• **Submitted:** 23.07.2024 • **Revised:** 03.02.2025 • **Accepted:** 26.03.2025

Abstract

Objectives. The development of contemporary models for the conversion of accents in foreign languages utilizes deep neural network architectures, as well as ensembles of neural networks for speech recognition and generation. However, restricted access to implementations of such models limits their application, study, and further development. Moreover, the use of these models is limited by their architectural features, which prevents flexible changes from being carried out in the timbre of the generated speech and requires the accumulation of context, leading to increased delays in generation, making these systems unsuitable for use in real-time multi-user communication scenarios. Therefore, the relevant task and aim of this work is the development of a method that generates native-sounding speech based on input accented speech material with minimal delays and the capability to preserve, clone, and modify the timbre of the speaker's voice.

Methods. Methods for modifying, training, and combining deep neural networks into a single end-to-end architecture for direct speech-to-speech conversion are applied. For training, original and modified open-source datasets were used.

Results. The work resulted in the development of a real-time accent conversion method with voice cloning based on a non-autoregressive neural network. The model comprises modules for accent and gender detection, speaker identification, speech conversion, spectrogram generation, and decoding the resulting spectrogram into an audio signal. As well as demonstrating high accent conversion quality while maintaining the original timbre, the short generation times of the applied method make it acceptable for use in real-time scenarios.

Conclusions. Testing of the developed method confirmed the effectiveness of the proposed non-autoregressive neural network architecture. The developed model demonstrated the ability to work in real-time information systems in English.

Keywords: accent conversion, speech synthesis, text-to-speech, voice conversion, machine learning, neural network

For citation: Nechaev V.A., Kosyakov S.V. Accent conversion method with real-time voice cloning based on a non-autoregressive neural network model. *Russian Technological Journal.* 2025;13(3):7–20. https://doi.org/10.32362/2500-316X-2025-13-3-7-20, https://www.elibrary.ru/PVYBDD

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели

В.А. Нечаев [@], **С.В. Косяков**

Ивановский государственный энергетический университет имени В.И. Ленина, Иваново, 153003 Россия
[@] *Автор для переписки, e-mail: nechaev@gapps.ispu.ru*

• **Поступила:** 23.07.2024 • **Доработана:** 03.02.2025 • **Принята к опубликованию:** 26.03.2025

Резюме

Цели. В настоящее время при разработке моделей для преобразования речи с акцентом в речь без акцента используются архитектуры глубоких нейросетей, а также ансамбли предобученных нейросетей для распознавания и генерации речи. При этом доступ к реализациям таких моделей является ограниченным, что затрудняет их применение, изучение и дальнейшее развитие. Также использование данных моделей ограничено особенностями архитектуры, которая не позволяет гибко менять тембр генерируемой речи и требует накопления контекста, что ведет к увеличению задержки при генерации и делает данные системы непригодными для использования в сценариях коммуникации двух и более людей в реальном времени. В связи с этим актуальной задачей и целью настоящей работы является разработка метода, позволяющего на основе входной речи с акцентом генерировать речь без акцента с минимальными задержками с возможностью сохранения, клонирования и модификации тембра говорящего, что позволит преодолеть ограничения текущих моделей.

Методы. Применены методы модификации, обучения и объединения глубоких нейросетей в единую сквозную архитектуру для прямого преобразования речи в речь. Для обучения использованы оригинальные и модифицированные наборы данных из открытых источников.

Результаты. Разработан метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели, которая состоит из модулей определения акцента и пола, идентификации говорящего, преобразования речи в фонетическое представление, генерации спектрограммы и декодирования полученной спектрограммы в аудиосигнал. Метод демонстрирует высокое качество конвертации акцента с сохранением оригинального тембра, а также низкие задержки при генерации, приемлемые для использования в сценариях реального времени.

Выводы. Апробация разработанного метода подтвердила эффективность предложенной неавторегрессионной нейросетевой архитектуры. Разработанная прикладная нейросетевая модель продемонстрировала возможность работы в информационных системах на английском языке в режиме реального времени.

Ключевые слова: конвертация акцента, генерация речи, распознавание речи, конвертация голоса, машинное обучение, нейронная сеть

Для цитирования: Нечаев В.А., Косяков С.В. Метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели. *Russian Technological Journal*. 2025;13(3):7–20. https://doi.org/10.32362/2500-316X-2025-13-3-7-20, https://www.elibrary.ru/PVYBDD

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

One of the most important channels of interaction between businesses and their customers is communication through voice communication. This is evidenced by the development and widespread use of call centers, which can now be established to operate on a cross-national and cross-regional basis. In such cases, English is very often used to overcome the interlingual barrier despite the operators and clients of call centers not all being native speakers of this language. As a result, situations may arise when a customer abandons a communication due to the difficulty of mutual understanding with the operator, which leads to economic losses. With the development of artificial intelligence systems, accent conversion software systems have been used to solve this problem, which enable to reduce the speaker's accent to a certain extent [1]. Such systems can also be used in the process of teaching foreign languages [2, 3], re-recording and enhancing the quality of previously recorded speech [4], and improving the quality of existing speech recognition systems [5]. Despite the considerable developments that have recently taken place in this area of research, the problem of improving and enhancing the quality of such systems remains relevant.

Accents in speech, representing an integral feature of pronunciation, can be divided into native accents, which depend on many regional and cultural factors, and foreign accents [6]. At the same time, foreign accent differs from native accent at the segmental (phonemes) and suprasegmental (intonation, accents, rhythmics) levels [7]. A foreign accent manifests itself when a native speaker of one language (L1 speech) speaks in another non-native or second language (L2 speech) [8]. L2 speech can be less intelligible to native speakers than L1 speech based on similar content [9], resulting in reduced comprehension of and trust in what is said, negative attitudes towards the speaker, and other forms of discrimination by native speakers [10–12].

Early methods for accent conversion in the generation phase are based on reference L1 examples corresponding to L2 speech [13–16]. For each L2 phrase, a corresponding L1 phrase is required. The practical application of such models is limited due to insufficient data to cover all possible speech variations. Such approaches also require significant resources for data collection and processing, which increases the development time and cost of such systems. In addition, the strict adherence to pairwise examples may reduce the versatility and scalability of the technology by limiting its ability to adapt to new accents or speech styles that were not included in the original dataset.

In subsequent developments, this limitation was overcome, meaning that reference examples are not required at the inference stage [17–22]. However, parallel datasets containing similar L2 and L1 phrases are still used for model training [17, 20], necessitating difficult and expensive operations for obtaining a sufficient amount of such data. Moreover, the autoregressive recurrent neural networks used by the described methods complicates the process of training them. Another method [19] uses pretrained neural networks to convert text to speech. To preserve individual voice characteristics, a separate model would need to be trained for each target speaker, making multi-user use difficult. Methods [21, 22] based on predicting the duration of each generated phoneme, which transforms the original L2 speech duration and speaker identity, require the accumulation of context, increasing generation time and complicating real-time use. Although the method described in [18] is not subject to the disadvantages listed above, the implementation of the model is limited to a finite set of accents, whose identifiers have to be determined during the model training phase. This complicates the process of training the data and applying the model with accents that have not been previously represented in it.

The present work set out to develop an accent conversion method that overcomes the above problems and shortcomings, capable of converting any speech from L2 to L1 without using reference examples or parallel data in the training and generation stages, which greatly simplifies, cheapens and speeds up the process of adapting the system to new accents.

1. RESEARCH OBJECTIVES

Speakers perceive literacy and fluency, on the one hand, and accent, on the other hand, as separate entities; improvement of one of them leads to an enhanced overall perception of L2 speech by native speakers [9, 23]. At the same time, absolutely accurate reproduction of L1 speech by a non-native speaker is difficult to achieve in practice due to differences in phonetic interference and speech perception by native speakers [24]. When solving the task of accent conversion, it is important to preserve the speaker's individual vocal features (timbre, pitch, loudness), i.e., it is necessary to perform voice cloning [25] while modifying segmental and suprasegmental characteristics associated with the foreign accent and pronunciation [13, 17, 18]. This is especially important in situations where it is necessary to preserve the emotional coloring, expressiveness, and individual features of speech, including voice features related to the speaker's gender.

In order to fulfill these conditions, it is necessary to use several interconnected, end-to-end architecture models for accent and gender detection, speaker embedding (SE), speech-to-phonetic (STP) representation, and spectrogram generation, as well as the decoding of the resulting spectrogram into an audio signal.

In addition, in real-time scenarios, such as human voice communication using high-speed communication channels, it is necessary to ensure minimal delays in the generation and transmission of processed speech [26–28]. Accent translation should be performed in real-time without the use of recurrent networks [29] to avoid the error accumulation effect associated with sequential output generation.

The training phase is based on publicly available open data for training the speech recognition and generation systems. In addition, the method set out to be independent of the availability of benchmark examples and parallel data at the training and inference stages.

2. RESEARCH METHODS

2.1. Architecture of the developed system

The developed accent conversion method includes several interrelated models integrated into a single end-to-end architecture for accent and gender detection, SE, STP conversion, spectrogram generation, and decoding of the obtained spectrogram into an audio signal. Figure 1 shows the general scheme of interaction of the above models at the output stage (generation of output L1 audio).

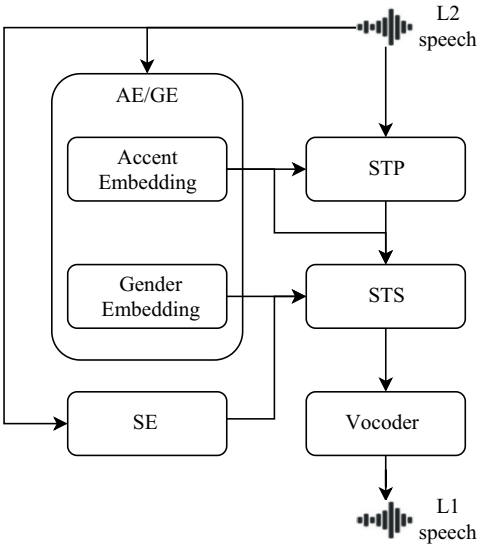

Fig. 1. General derivation scheme of the accent conversion method with voice cloning

The L2 speech audio signal is fed to the input of the STP model, to the input of the Accent and Gender Embedding (AE/GE) model, and to the input of the SE model. The accent embedding affects the generation of the phonetic representation, which in vectorized form is fed to the input of the speech-to-speech (STS) and mel spectrogram generation model. The AE/GE vector representations (embeddings) are also fed to the input of the STP model, as well as the output of the SE model, which is a vector representation of individual voice characteristics (timbre). The resulting spectrogram is converted into an L1 speech audio signal using a decoding vocoder model.

The overall pipeline of L1 speech generation from the original L2 speech can be simplistically represented as a formula:

$$a_{L1} = F_V(F_{STS}(F_{STP}(a_{L2}, F_{AE}(a_{L2})), F_{AE}(a_{L2}), F_{GE}(a_{L2}), F_{SE}(a_{L2}))), \quad (1)$$

where a_{L1} is the generated L1 speech audio signal; a_{L2} is the input L2 speech audio signal; F_V is the vocoder model; F_{STS} is the STS model; F_{STP} is the STP model; F_{AE} is AE in the AE/GE model; F_{GE} is GE in the AE/GE model; F_{SE} is the SE model, vector representation of individual voice characteristics.

In order to obtain a single end-to-end model of accent conversion, it is necessary to perform the training process of each model sequentially. Thus, the AE/GE and SE models are independent of other models and their training can be performed in any order. The output of the trained AE/GE model will be required at the stage of obtaining the STP model. All previous models (AE/GE, SE, STP) are required to derive the STS model. The training of the vocoder model is based on the output of the STS model.

2.2. AE/GE model

In order to obtain fixed length vectors representing the accent and gender features of the speaker, the model is first trained for the classification task. In such a configuration, class labels used in the training process are returned by the model at the output of the last layer, while vector representations used as voice features are taken from a special intermediate layer.

This and other models use a preprocessor based on fast Fourier transform, which converts the incoming audio signal (time domain) into a mel spectrogram (frequency domain), showing the frequency content of the audio signal on a perceptual mel scale, which approximates the nonlinear frequency response of the human ear. Here, the sampling frequency (sampling rate) is 22050 Hz and the window width is 1024 sound fragments (samples), while the window step is 256 samples and the number of generated mel bands is 80.

Figure 2 shows the training scheme of the accent and gender detection model. It contains blocks of convolutional network of Jasper architecture of 3 × 3 configuration [30]. The accent decoder and gender decoder, which have a common architecture, consist of an attention pooling layer [31], a normalization layer, a convolutional layer to obtain AE and GE of dimension 192, and a linear layer to obtain (predict) the accent class (AC) and gender class (GC).

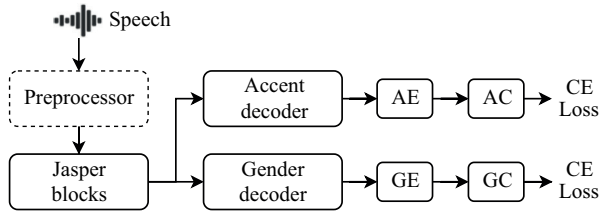

Fig. 2. Training diagram of the AE/GE model. CE are cross entropies

After feeding the audio signal to the preprocessor, the mel spectrogram is fed to the Jasper blocks, as well as, in parallel, to the accent decoder and the gender decoder, along with the corresponding fully connected layers in the output to obtain the accent and gender prediction vectors. During the model training process, the sum of CEs is minimized:

$$L_{\text{AE,GE}} = (x_{\text{a}}, y_{\text{a}}, x_{\text{g}}, y_{\text{g}}) = \\ = -\sum_{i=1}^{A} y_{\text{a}_i} \ln\left(\frac{\exp x_{\text{a}_i}}{\sum_{k=1}^{A} \exp x_{\text{a}_k}}\right) - \sum_{j=1}^{G} y_{\text{g}_j} \ln\left(\frac{\exp x_{\text{g}_j}}{\sum_{l=1}^{G} \exp x_{\text{g}_l}}\right), \quad (2)$$

where $L_{\text{AE, GE}}$ is the overall loss function of the AE/GE model; A is the number of ACs (40); G is the number of GCs (2); x_{a} are accent predictions; x_{g} are gender predictions; y_{a} are ground truth accent labels; y_{g} are ground truth gender labels.

2.3. SE model

Figure 3 shows the training scheme of the SE model and tone vector representation. This scheme contains an input convolutional neural network of SincNet architecture [32], layers of X-Vector DNN[1] model [33], and a layer for obtaining vector representations of dimensionality 512.

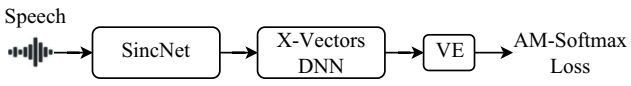

Fig. 3. Diagram of the SE model training. VE is the voice embedding—vector representation of the speaker's individual voice characteristics

Unlike the AE/GE model, the audio signal is not pretransformed into a mel spectrogram, i.e., the digitized audio signal in the time domain with a sampling rate of 16000 Hz is fed to the band-pass filters of the SincNet architecture, then to the convolution layers of the X-Vector DNN and the output fully connected layer, which provides a vector representation of individual voice characteristics (timbre) at the output. In the process of model training, the problem of representation learning is solved while minimizing the Additive Angular Margin (AAM Loss) function [34].

2.4. STP conversion model

The next step is speech recognition taking into account the speaker's accent. For this purpose, it is necessary to obtain a model for converting speech into phonetic or textual representation. The training scheme of the STP model is shown in Fig. 4. The dotted line represents the blocks that are fixed (frozen) during the backpropagation process, i.e., the weights in these blocks are not updated, but their previously obtained states are used.

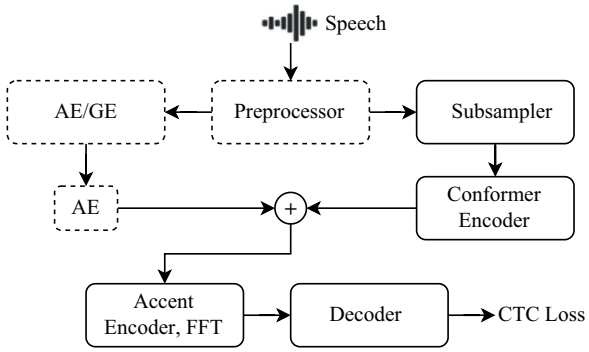

Fig. 4. Diagram of STP model training

The speech audio signal is fed to the input of the preprocessor as previously described and then in parallel to the AE/GE model to obtain the AE and to the convolutional dimensionality reduction block (Subsampler) with a factor of 4. Further conversion is carried out using the Conformer encoder, which comprises a 12 module Conformer architecture [35] having an internal dimensionality of 512 consisting of fully connected [36], convolutional [37], and attention mechanism transformer layers [38]. AE is then normalized, reduced to dimensionality 512, summed with the output of the Conformer encoder, and fed to the input of the accent encoder, which has a single stack feed-forward transformer (FFT) architecture [39]. The output of the accent encoder is further utilized in the STS model as a distribution of phonetic tokens. Finally, the output signal of the accent encoder is fed to the decoder, which has a single-layer convolutional architecture with Softmax activation function, and forms at the output a vector of predictions of textual tokens of dimension equal to the size of the tokenizer dictionary (128) plus one (for a blank token). During model training, the Connectionist Temporal Classification (CTC) Loss function [40] is minimized, which calculates the loss between the continuous (unsegmented) time series and the target sequence:

[1] Deep neural network.

$$L_{\text{STP}}(x,y) = -\ln\left(\sum_{\rho\in A_{x,y}}\prod_{t=1}^{T} x_{\rho_t}\right), \quad (3)$$

where L_{STP} is the loss function of the STP model (CTC Loss); x is the probabilities of text tokens predicted by the model; y is the sequence of text tokens from the target text; ρ is the alignment path x predictions to reduce to y sequence by removing all blank tokens and merging repeated tokens; $A_{x,y}$ is the set of all possible alignment paths; T is the number of predicted tokens in x; x_{ρ_t} is the probability of a particular predicted token at step t given the chosen alignment path ρ.

2.5. STS conversion and spectrogram generation model

The previous models are combined into a single architecture for STS conversion and spectrogram generation. Figure 5 presents a schematic of its training. The STS model includes the previously discussed blocks of preprocessor, accent, gender (AE/GE model) and speaker's timbre (SE model) with the corresponding modules of vector representations (AE, GE, VE), as well as the STP conversion block (STP model). All these blocks are marked with a dotted line due to their training was performed earlier and is not performed at the stage of STP-model training. Moreover, an untrained block based on normalized cross correlation function and median smoothing was added to the architecture to extract the fundamental or lowest frequency of the periodic sound signal (F0), which is perceived by the human ear as pitch [41, 42].

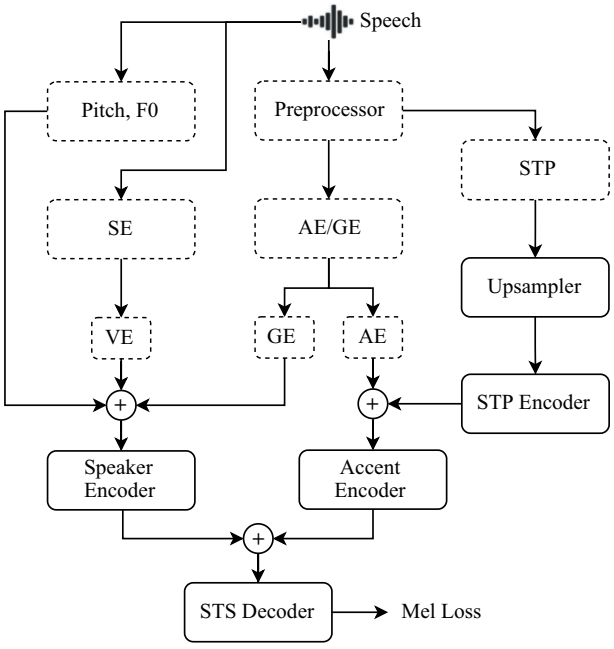

Fig. 5. STS Model training diagram

The speech audio signal is fed to the preprocessor, pitch block and SE model input. The mel spectrogram from the preprocessor is fed to the inputs of the AE/GE and STP models. The phonetic representation from the STP model is fed to an upsampler with a factor of 4 to equalize the original and generated spectrograms, consisting of two convolutional 1D-transposed layers and two rectified linear unit (ReLU) activation functions placed after each convolutional layer. After the upsampler, the phonetic representation is transformed using a STP encoder, which has a six-stack feed-forward transformer (FFT) architecture [39] used in the Fastpitch architecture as an input unit operating in the token domain [43], with inner and outer dimensions of 1536 and 384, respectively. The vector representations of accent, gender, speaker's timbre and pitch profile are normalized and reduced to dimensionality 384. Next, the accent vectors and the output of the STP encoder are summed and fed to the input of the accent encoder (1 stack FFT). Similarly, the pitch, timbre, gender vectors are summed and fed to the input of the speaker encoder (1 stack FFT). Thus, the speaker encoder aggregates speech properties related to individual voice characteristics except accent, which in turn is the responsibility of the accent encoder. The sum of the output vectors of the speaker encoder and the accent encoder is fed to the input of a STS decoder consisting of 6 stacks of Fastpitch architecture FFTs from the output mel area [43]. Finally, the vector is projected to dimension 80 to match the original number of mel areas. During the training process, the loss function is minimized based on the standard deviation:

$$L_{\text{STS}}(x,y) = \frac{1}{\sum_{i=1}^{N} d_i}\sum_{i=1}^{N} d_i (y_i - x_i)^2, \quad (4)$$

where L_{STS} is the loss function of the STS model (Mel Loss); N is the number of elements in the mel spectrogram; x is the mel spectrogram predicted by the model; y is the target mel spectrogram; d is the mask of the spectrogram duration for collection into a batch of fixed size, consisting of values 1 (the element should be considered) and 0 (the element should not be considered), obtained from the duration of the predicted spectrogram.

2.6. Model of sound signal generation from mel spectrogram (vocoder)

Mel spectrogram of L1 speech in the frequency domain obtained using the STS model is converted into a sound signal in the time domain. For this purpose, a model based on generative-adversarial networks (HiFi-GAN) [44] is used. The output audio signal has a sampling rate of 22050 Hz. The model is

trained as follows: the audio signal from the training dataset is converted into a mel spectrogram using the STS model, then the resulting spectrogram is passed to a vocoder and converted into an audio signal. Using the received and original audio signals, the loss functions for the generator and discriminator are calculated as described in [44].

2.7. STS conversion and spectrogram generation simplified model (Ablation)

In order to conduct comparative experiments, a simplified version of the accent conversion model was also developed, the schematic of which is shown in Fig. 6.

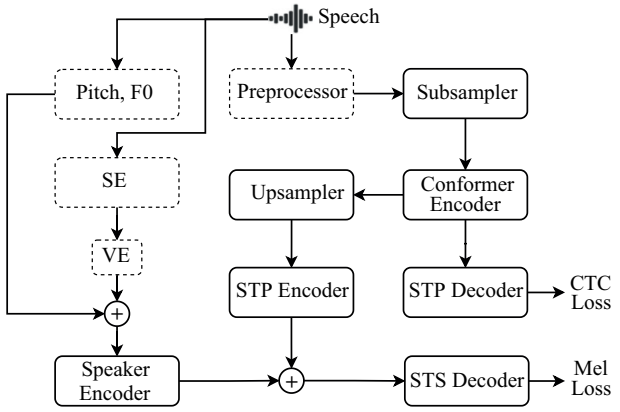

Fig. 6. Schematic of a simplified accent conversion model

This simplified model excludes the AE/GE model, as well as all related encoders in the STP and STS models. Thus, in the resulting simplified model, the output is not conditioned on accent and gender properties. In addition, the training of the STP model was conducted not separately, but simultaneously with the STS model without fixing the weights of the STP with minimization of the sum of the CTC Loss and Mel Loss functions.

3. PRACTICAL APPLICATION OF THE METHOD

3.1. Model training

AE/GE model was trained on the following datasets: CMU-ARCTIC [45], L2 ARCTIC [46], Speech Accent Archive [47], Common Voice 16.1 [48]. All of them represent audio recordings of speech in English, their corresponding textual transcriptions, and contain additional meta-information about accent, gender and, in some cases, native language, place of residence and age of the speaker. Using this information, the audio files were grouped into 40 classes denoting native or foreign English accents, e.g., British, American, Russian, Indian and South Asian, Canadian, German, Australian, African, Japanese, Eastern European, etc. The gender of the speaker is also highlighted. The total duration of the audio files marked in this way amounted to 1087.6 h for the training sample, and 7.6 h for the validation and test samples.

VoxCeleb1 [49] and VoxCeleb2 [50] data collections with a total duration of 2794 h have been used to train the SE model. These sets represent grouped speech audio recordings of 7363 individuals. Audio recordings pertaining to one person are presented during training as positive examples and, conversely, those pertaining to different people as negative examples.

STP model was trained on data from CMU-ARCTIC [45], L2 ARCTIC [46], Common Voice 16.1 [48], LibriSpeech [51], NPTEL2020[2], VCTK [52], GigaSpeech [53]. The mentioned sets consist of audio recordings of English speech with different accents and corresponding text transcriptions. The total duration of the pooled training sample was 6107 h and the validation sample was 48 h. The text transcriptions were normalized, i.e., converted from the canonical written form to the spoken form [54], which is especially important for numbers and abbreviations, and were also brought to a unified form: lower case into ASCII format, punctuation, special characters and additional indentation were removed. A SentencePiece tokenizer [55] with a dictionary size of 128 was trained on the training part of the texts, with which all the texts are processed during the training and evaluation of the model.

STS model and vocoder were trained using the following datasets: CMU-ARCTIC [45], L2 ARCTIC [46], VCTK [52], LibriTTS-R [56], LJ Speech[3]. When we split the data into training and validation samples, their duration was 681 and 17.6 h, respectively. Only audio information without textual markup is used in the training process.

A simplified model (ablation) was trained on data for the STP and STS models.

In order to train, evaluate, and use the described models, code has been developed using the open-source libraries Pytorch [57] and NVIDIA NeMo [58]. The implementation and weights of the vector representation model of the speaker's timbre (SE model) are taken from the Pyannote library [59]. Training was conducted on a server with 8 NVIDIA Tesla V100 graphics processing units (GPUs).

AE/GE model was trained using the stochastic gradient descent optimizer with a learning rate of $1 \cdot 10^{-3}$, weight decay $2 \cdot 10^{-4}$, momentum 0.9, and a Cosine

[2] NPTEL2020 – Indian English Speech Dataset. https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset. Accessed May 01, 2024.

[3] Ito K., Johnson L. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. Accessed May 01, 2024.

Annealing scheduler for 200 epochs. To train the STP and STS models, the AdamW optimizer was used with a learning rate of $1 \cdot 10^{-3}$, a regularization factor of 0.001, and a similar scheduler as for the AE/GE model for 50 epochs for each model. Fine-tuning of the HiFi-GAN vocoder model was performed by initializing model weights obtained from open sources [44], using the AdamW optimizer and a learning rate of $1 \cdot 10^{-6}$ for 40 epochs. The training of the simplified model (ablation) was carried out with similar parameters used in the STS model.

Table 1 shows the number of trainable parameters of the models optimized during training. In total, the considered accent conversion architecture (full STS), consisting of several interconnected models, has 164 mln parameters.

Table 1. Number of trained parameters

Model	Number of parameters, mln
AE/GE	24.9
SE	4.3
STP	82.1
STS	52.7
Full STS	**164**
Vocoder	84.7
Total	248.7

3.2. Performance assessment

The model performance was assessed on a Linux server running a single NVIDIA Tesla T4 GPU, an 8-core virtual central processing unit (vCPU), and 16 GB of random-access memory (RAM). To accomplish this, the model was first exported to the open source ONNX format and then deployed using NVIDIA Triton open-source software. Using the program interface of the NVIDIA Triton-deployed model and a 5 s test audio file containing English L2 speech, we measured the response generation latency at 200 iterations. As a result, the average latency was 52 ms and throughput was 96 RTFX.

Performance assessment results of the accent conversion model show low generation delays. Together with the features of the architecture, which does not require the accumulation of long context but can handle segments of less than 0.25 s duration, this makes it possible to apply the proposed model in real-time dialog when response delays affect communication [26–28].

3.3. Objective quality assessment

Open-source data as well as pretrained speech recognition models were used to perform objective quality assessment. Using the proposed accent conversion method, an audio file was generated for each example from the test set. Quality metrics were then calculated for the original and corrected audio files. Table 2 presents the results of the objective quality assessment.

As test datasets, we used subsamples totaling 26.9 h that did not participate in the process of training the accent conversion model and its components. All of them include text transcriptions and audio files of English speech with different native and non-native accents from open sources:

- 3.2 h from CMU-ARCTIC [45], L2 ARCTIC [46] (ARCTIC), 10 accents: American, English, Chinese, Indian, Korean, Vietnamese, Spanish, Arabic, Dutch, German;
- 3.1 h from Common Voice [48], 12 accents: American, English, Indian, Australian, African, Chinese, Filipino, Malaysian, German, Russian, French, Eastern European;
- 15.2 h from NPTEL2020, Indian accent;
- 5.4 h of Afrispeech-200 [60], African accent (Yoruba, Swahili, Igbo, Zulu, Tswana, Idoma, Afrikaans).

Speech recognition models obtained from open sources were used: Conformer [35], Citrinet [61], and Whisper [62]. In this case, the Whisper model is taken in two variants: large multilingual (L. Mult.) and medium English (M. En.). Recognition was performed on audio files without processing and on audio files after accent conversion. The recognized and true transcriptions were then reduced to a single form using normalization [54], after which quality metrics were compared and counted: word error rate (WER), character error rate (CER). In Table 2, the best results for each pair: the test dataset and the speech recognition model are highlighted in bold type.

As can be seen from the results, in almost all cases the accent conversion method improves the recognition of the pretrained models, as indicated by the reduced values of word and character error rates. The accent conversion model improves speech quality by making it more recognizable.

3.4. Subjective quality assessment

Group-based listening tests were conducted with 53 participants from different countries with an English language proficiency level of at least B2 according to the CEFR[4] scale. For this purpose,

[4] CEFR (Common European Framework of Reference) is the system of foreign language proficiency levels used in Europe. https://www.coe.int/en/web/common-european-framework-reference-languages. Accessed May 01, 2024.

Table 2. Results of accent conversion model assessment with speech recognition models. Data after conversion are marked as 'conv.'

Test dataset	Speech recognition model			
	Conformer	Citrinet	Whisper L. Mult.	Whisper M. En.
WER, %				
ARCTIC	9.57	11.73	16.23	8.91
ARCTIC conv.	**8.78**	**11.55**	**12.69**	**8.68**
Common Voice	**9.07**	25.80	36.89	11.26
Common Voice conv.	9.12	**23.38**	**22.71**	**10.62**
NPTEL2020	29.18	29.88	16.41	15.18
NPTEL2020 conv.	**25.26**	**29.41**	**13.87**	**11.64**
Afrispeech-200	43.2	46.24	37.91	33.61
Afrispeech-200 conv.	**35.19**	**39.49**	**35.56**	**29.96**
CER, %				
ARCTIC	3.73	4.85	10.30	3.98
ARCTIC conv.	**3.52**	**4.68**	**6.06**	**3.92**
Common Voice	**3.75**	8.74	21.41	5.66
Common Voice conv.	3.77	**8.29**	**13.63**	**5.22**
NPTEL2020	16.87	17.70	11.94	10.67
NPTEL2020 conv.	**14.79**	**17.01**	**10.10**	**9.44**
Afrispeech-200	31.52	34.79	24.30	20.04
Afrispeech-200 conv.	**27.86**	**28.92**	**23.15**	**18.88**

each of the participants was given instructions, where within each experiment they were asked to listen to 1 or 2 audio files and give their assessment of compliance with the quality criterion on a five-point scale, where '1' – definitely does not comply, '2' – rather does not comply, '3' – compromise, '4' – rather complies, '5' – exactly complies. The resulting scores were then used to calculate the mean opinion score (MOS) for each experiment. The results are presented in Table 3.

Twenty pairs of audio files from test subsamples of L2 ARCTIC [46] and NPTEL2020 datasets with non-native English accent (Original) were randomly selected as audio samples: Indian, Chinese, Korean, Vietnamese, Spanish, Arabic, and German. Each original audio pair represents a recording of the same speaker. For each selected audio file (40 in total), variants were generated using a simplified accent conversion model (Ablation) and using the proposed model (Proposed). A total of 3 experiments were conducted to evaluate voice naturalness, speaker similarity and absence of foreign accent. In all experiments, at least 3 evaluations were asked for each type of sound sample. The test samples themselves were varied across experiments, eliminating repetition. The sample could be listened to an unlimited number of times before scoring. Thus, each interviewee made a total of 9 to 12 evaluations.

Table 3. Results of subjective quality assessment (MOS with 95% confidence interval)

Examples	Naturalness of voice	Similarity of the speakers	Absence of foreign accent
Original	4.83 ± 0.10	4.91 ± 0.08	2.06 ± 0.18
Ablation	3.38 ± 0.13	3.92 ± 0.15	3.58 ± 0.17
Proposed	4.04 ± 0.16	4.30 ± 0.18	4.11 ± 0.14

When assessing the naturalness of the voice, participants were asked to determine on a five-point scale how natural the speech in the audio example sounds, i.e., whether the listener gets the impression that it is a real live human voice and not a generated or robotized speech. A score of '1' means that the voice is definitely artificial, synthesized using computer-generated methods, and '5' means that the example sounds like speech produced using analog or digital sound recording methods of a real human voice. Interviewees were also advised not to pay attention to the presence or absence of background noise in the recording in order to concentrate on speech evaluation.

In order to conduct an experiment on speaker similarity evaluation, pairs of audio recordings were prepared: Original–Original, Original–Ablation, and Original–Proposed. Meanwhile, the first pair includes recordings from the original data only, which are recordings of the same speaker but uttering different phrases. The other pairs include an original recording of one phrase and a generated version of another phrase by the same speaker. Participants were asked to listen to such pairs of audio recordings and decide whether they were spoken by the same person, i.e., how similar the timbre in one file is to the timbre in the other file. Score '1' is the speech in the audio recordings definitely belongs to different people, '5' is the timbre of the speakers in the audio recordings is identical, belonging to one person. Interviewees were recommended to ignore the L1 and L2 accent properties during the evaluation in order to focus on comparing the overtone coloration of the voice.

To assess the absence of a foreign accent, participants were asked to listen to an English-language audio file and decide how much of a foreign accent they thought the recording contained. English and American accents were assumed to be native L1 and all other accents were assumed to be non-native L2. A score of '1' means that the speech has a pronounced foreign L2 accent, '5' means that the speech is definitely an English-speaking L1 without a foreign accent.

The analysis of the table shows that the highest estimates of voice naturalness and speaker similarity show the original examples, which is obvious, since they are obtained without using speech synthesis methods, and together with the lowest estimate of the absence of foreign accent demonstrates the calibration of the opinions of the experiment participants on real data. Adding the AE/GE model to the overall scheme of the accent conversion model significantly improves the quality of generation, this is demonstrated by the improved results compared to the simplified Ablation model. In all subjective experiments, the proposed model shows a score higher than '4', meaning that, in the opinion of the interviewees, the model rather meets the specified quality criteria.

CONCLUSIONS

The study presents an accent conversion method that converts any L2 speech with a pronounced foreign accent into L1 speech, does not depend on the availability of reference examples and parallel data at the training and generation stages, which greatly simplifies, cheapens and speeds up the process of adapting the system to new accents.

The proposed non-autoregressive model, which does not use recurrent networks in its architecture, features an accelerated training process and real-time accent translation while avoiding the error accumulation effect associated with sequential output generation.

The described method also includes an algorithm for cloning the speech characteristics of the speaker to preserve his or her vocal identity even following accent conversion. This is especially important in situations where emotional coloration, expressiveness, and individual speech features are required. In addition, the method enables real-time modification of voice characteristics such as accent, timbre, and gender-related voice features during the generation process by copying the corresponding characteristics from an audio sample, making it applicable in a wider range of scenarios than previous developments.

The model demonstrates high quality accent conversion while preserving the original timbre, as well as low generation latency acceptable for use in real-time scenarios.

The method can be used to:

1) convert English-speaking speech with a foreign L2 accent into L1 speech without a foreign accent;
2) improve speech quality and, as a consequence, to improve the recognition quality of existing systems;
3) copy and change the speaker's voice characteristics in real time;
4) apply real-time accent conversion in a dialog mode.

The developed applied neural network model demonstrated the ability to work in real-time English language information systems. The results of the study can be applied to the development of voice modification systems, as well as speech recognition and generation systems.

Authors' contribution
All authors equally contributed to the research work.

REFERENCES

1. McMillin D.C. Outsourcing identities: Call centres and cultural transformation in India. *Economic and Political Weekly*. 2006;41(3):235–241.
2. Felps D., Bortfeld H., Gutierrez-Osuna R. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*. 2009;51(10):920–932. https://doi.org/10.1016/j.specom.2008.11.004
3. Probst K., Ke Y., Eskenazi M. Enhancing foreign language tutors–In search of the golden speaker. *Speech Communication*. 2002;37(3–4):161–173. https://doi.org/10.1016/S0167-6393(01)00009-7
4. Türk O., Arslan L. M. Subband based voice conversion. In: *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH 2002*. *Interspeech*. 2002. P. 289–292.
5. Biadsy F., Weis, R.J., Moreno P.J., Kanevsky D., Jia Y. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *Interspeech*. 2019. P. 4115–4119. http://doi.org/10.21437/Interspeech.2019-1789
6. Birner B. *Why Do Some People Have an Accent?* Linguistic Society of America. Washington, DC. 1999. 6 p.
7. Baese-Berk M.M., Morrill T.H. Speaking rate consistency in native and non-native speakers of English. *J. Acoust. Soc. Am*. 2015;138(3):EL223–EL228. https://doi.org/10.1121/1.4929622
8. Piske T., MacKay I.R.A., Flege J.E. Factors affecting degree of foreign accent in an L2: A review. *J. Phonetics.* 2001;29(2):191–215. https://doi.org/10.1006/jpho.2001.0134
9. Munro M.J., Derwing T.M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*. 1995;45(1):73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x
10. Lev-Ari S., Keysar B. Why don't we believe non-native speakers? The influence of accent on credibility. *J. Exp. Soc. Psychol*. 2010;46(6):1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025
11. Rubin D.L., Smith K.A. Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *Int. J. Intercult. Relat*. 1990;14(3):337–353. https://doi.org/10.1016/0147-1767(90)90019-S
12. Nelson Jr. L.R., Signorella M.L., Botti K.G. Accent, gender, and perceived competence. *Hispanic J. Behavior. Sci.* 2016;38(2):166–185. https://doi.org/10.1177/0739986316632319
13. Zhao G., Gutierrez-Osuna R. Using phonetic posteriorgram based frame pairing for segmental accent conversion. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2019;27(10):1649–1660. https://doi.org/10.1109/TASLP.2019.2926754
14. Zhao G., Sonsaat S., Levis J., Chukharev-Hudilainen E., Gutierrez-Osuna R. Accent conversion using phonetic posteriorgrams. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2018. P. 5314–5318. https://doi.org/10.1109/ICASSP.2018.8462258
15. Aryal S., Gutierrez-Osuna R. Can voice conversion be used to reduce non-native accents? In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2014. P. 7879–7883. https://doi.org/10.1109/ICASSP.2014.6855134
16. Ding S., Zhao G., Gutierrez-Osuna R. Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*. 2022;72:101302. https://doi.org/10.1016/j.csl.2021.101302
17. Quamer W., Das A., Levis J., Chukharev-Hudilainen E., Gutierrez-Osuna R. Zero-shot foreign accent conversion without a native reference. *Proc. Interspeech*. 2022. http://doi.org/10.21437/Interspeech.2022-10664
18. Jin M., Serai P., Wu J., Tjandra A., Manohar V., He Q. Voice-preserving zero-shot multiple accent conversion. In: *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2023. P. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094737
19. Zhou Y., Wu Z., Zhang M., Tian X., Li H. TTS-guided training for accent conversion without parallel data. *IEEE Signal Proc. Lett*. 2023;30:533–537. https://doi.org/10.1109/lsp.2023.3270079
20. Zhao G., Ding S., Gutierrez-Osuna R. Converting foreign accent speech without a reference. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021;29:2367–2381. https://doi.org/10.1109/TASLP.2021.3060813
21. Liu S., Wang D., Cao Y., Sun L., Wu X., Kang S., Wu Z., Liu X., Su D., Yu D., Meng H. End-to-end accent conversion without using native utterances. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2020. P. 6289–6293.

22. Zhou X., Zhang M., Zhou Y., Wu Z., Li H. Accented text-to-speech synthesis with limited data. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2024;32:1699–1711. https://doi.org/10.1109/TASLP.2024.3363414
23. Pinget A.F., Bosker H.R., Quené H., De Jong, N.H. Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*. 2014;31(3):349–365. https://doi.org/10.1177/0265532214526177
24. Barkhudarova E.L. Methodological Problems in Analyzing Foreign Accents in Russian Speech. *Vestnik Moskovskogo universiteta. Seriya 9. Filologiya = Lomonosov Philology J.* 2012;6:57–70 (in Russ.).
25. Arik S., Chen J., Peng K., Ping W., Zhou Y. Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems* (*NeurIPS 2018*). 2018;31.
26. Cohen D. Issues in transnet packetized voice communication. In: *Proceedings of the fifth Symposium on Data Communications* (*SIGCOMM'77*). 1977. P. 6.10–6.13. https://doi.org/10.1145/800103.803349
27. Liang Y.J., Farber N., Girod B. Adaptive playout scheduling and loss concealment for voice communication over IP networks. *IEEE Trans. Multimedia*. 2003;5(4):532–543. https://doi.org/10.1109/TMM.2003.819095
28. Matzinger T., Pleyer M., Żywiczyński P. Pause Length and Differences in Cognitive State Attribution in Native and Non-Native Speakers. *Languages*. 2023;8(1):26. http://doi.org/10.3390/languages8010026
29. Medsker L.R., Jain L. (Eds.). *Recurrent Neural Networks. Design and Applications*. Boca Raton: CRC Press; 2001. 416 p.
30. Li J., Lavrukhin V., Ginsburg B., Leary R., Kuchaiev O., Cohen J.M., Nguyen H., Gadde R.T. Jasper: An End-to-End Convolutional Neural Acoustic Model. *Interspeech 2019*. 2019. https://doi.org/10.21437/interspeech.2019-1819
31. Dawalatabad N., Ravanelli M., Grondin F., Thienpondt J., Desplanques B., Na H. *ECAPA-TDNN Embeddings for Speaker Diarization.* arXiv preprint arXiv:2104.01466. 2021. https://doi.org/10.48550/arXiv.2104.01466
32. Ravanelli M., Bengio Y. Speaker recognition from raw waveform with SincNet. In: *2018 IEEE Spoken Language Technology Workshop* (*SLT*). IEEE; 2018. P. 1021–1028. https://doi.org/10.1109/SLT.2018.8639585
33. Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. X-vectors: Robust DNN embeddings for speaker recognition. In: *2018 IEEE international Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2018. P. 5329–5333. http://doi.org/10.1109/ICASSP.2018.8461375
34. Deng J., Guo J., Xue N., Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2019. P. 4690–4699. https://doi.org/10.1109/CVPR.2019.00482
35. Gulati A., Qin J., Chiu C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented trans-former for speech recognition. *Proc. Interspeech 2020*. 2020. P. 5036–5040. https://doi.org/10.21437/interspeech.2020-3015
36. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*. 2010. P. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html
37. Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Cai J., Chen T. Recent advances in convolutional neural networks. *Pattern Recognition*. 2018;77:354–377. https://doi.org/10.1016/j.patcog.2017.10.013
38. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017;30:5999–6009. https://doi.org/10.48550/arXiv.1706.03762
39. Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.Y. FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*. 2019;32. https://doi.org/10.48550/arXiv.1905.09263
40. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning*. 2006. P. 369–376. https://doi.org/10.1145/1143844.1143891
41. Ghahremani P., BabaAli B., Povey D., Riedhammer K., Trmal J., Khudanpur S. A pitch extraction algorithm tuned for automatic speech recognition. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2014. P. 2494–2498. http://doi.org/10.1109/ICASSP.2014.6854049
42. Gerhard D. *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Masters Thesis. Regina, SK, Canada: Department of Computer Science, University of Regina; 2003. 23 p.
43. Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2021. P. 6588–6592. https://doi.org/10.1109/ICASSP39728.2021.9413889
44. Kong J., Kim J., Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*. 2020;33:17022–17033. http://doi.org/10.48550/arXiv.2010.05646
45. Kominek J., Black A.W. The CMU Arctic speech databases. In: *Fifth ISCA Workshop on Speech Synthesis*. 2004. P. 223–224.
46. Zhao G., Sonsaat S., Silpachai A., Lucic I., Chukharev-Hudilainen E., Levis J., Gutierrez-Osuna R. L2-ARCTIC: A Non-native English Speech Corpus. *Interspeech 2018*. 2018. P. 2783–2787. http://doi.org/10.21437/Interspeech.2018-1110
47. Weinberger S.H., Kunath S.A. The Speech Accent Archive: towards a typology of English accents. In: *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Brill; 2011. P. 265–281. https://doi.org/10.1163/9789401206884_014
48. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Henretty M., Morais R., Saunders L., Tyers F., Weber G. Common Voice: A Massively-Multilingual Speech Corpus. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020. P. 4218–4222. https://doi.org/10.48550/arXiv.1912.06670

49. Nagrani A., Chung J.S., Zisserman A. Voxceleb: a large-scale speaker identification dataset. *Interspeech 2017*. 2017. http://doi.org/10.21437/Interspeech.2017-950
50. Chung J., Nagrani A., Zisserman A. VoxCeleb2: Deep speaker recognition. *Interspeech 2018*. 2018. http://doi.org/10.21437/Interspeech.2018-1929
51. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2015. P. 5206–5210. http://doi.org/10.1109/ICASSP.2015.7178964
52. Veaux C., Yamagishi J., MacDonald K. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. *University of Edinburgh. The Center for Speech Technology Research* (*CSTR*). 2017. https://doi.org/10.7488/ds/2645
53. Chen G., Chai S., Wang G., Du J., Zhang W., Weng C., Su D., Povey D., Trmal J., Zhang J., Jin M., Khudanpur S., Watanabe S., Zhao S., Zou W., Li X., Yao X., Wang Y., Wang Y., You Z., Yan Z. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In: *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*. International Speech Communication Association; 2021. P. 4376–4380. https://doi.org/10.21437/Interspeech.2021-1965
54. Bakhturina E., Zhang Y., Ginsburg B. Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization. *Proc. Interspeech 2022*. 2022. http://doi.org/10.48550/arXiv.2203.15917
55. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018. P. 66–71. https://doi.org/10.48550/arXiv.1808.06226
56. Koizumi Y., Zen H., Karita S., Ding Y., Yatabe K., Morioka N., Bacchiani M., Zhang Y., Han W., Bapna A. Libritts-r: A Restored Multi-Speaker Text-to-Speech Corpus. *arXiv preprint arXiv:2305.18802*. 2023. https://doi.org/10.48550/arXiv.2305.18802
57. Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. 2019;32: 8024–8035.
58. Kuchaiev O., Li J., Nguyen H., Hrinchuk O., Leary R., Ginsburg B., Kriman S., Beliaev S., Lavrukhin V., Cook J., Castonguay P., Popova M., Huang J., Cohen J. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*. 2019. https://doi.org/10.48550/arXiv.1909.09577
59. Bredin H., Yin R., Coria J.M., Gelly G., Korshunov P., Lavechin M., Fustes D., Titeux H., Bouaziz W., Gill M.P. Pyannote. Audio: neural building blocks for speaker diarization. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2020. P. 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974
60. Olatunji T., Afonja T., Yadavalli A., Emezue C.C., Singh S., Dossou B.F., Osuchukwu J., Osei S., Tonja A.L., Etori N., Mbataku C. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*. 2023;11:1669–1685. https://doi.org/10.1162/tacl_a_00627
61. Majumdar S., Balam J., Hrinchuk O., Lavrukhin V., Noroozi V., Ginsburg B. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv preprint arXiv:2104.01721*. 2021. http://doi.org/10.48550/arXiv.2104.01721
62. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR 202. 2023. P. 28492–28518. http://doi.org/10.48550/arXiv.2212.04356

About the Authors

Vladimir A. Nechaev, Teacher-Researcher, Ivanovo State Power Engineering University (34, Rabfakovskaya ul., Ivanovo, 153003 Russia). E-mail: nechaev@gapps.ispu.ru. RSCI SPIN-code 7002-3878, https://orcid.org/0009-0007-1449-3968

Sergey V. Kosyakov, Dr. Sci. (Eng.), Professor, Head of the Department of Computer Systems Software, Ivanovo State Power Engineering University (34, Rabfakovskaya ul., Ivanovo, 153003 Russia). E-mail: ksv@ispu.ru. Scopus Author ID 6507182528, ResearcherID H-5686-2018, RSCI SPIN-code 1371-9929, https://orcid.org/0000-0003-0231-0750

Об авторах

Нечаев Владимир Алексеевич, преподаватель-исследователь, ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина» (153003, Россия, Иваново, ул. Рабфаковская, д. 34). E-mail: nechaev@gapps.ispu.ru. SPIN-код РИНЦ 7002-3878, https://orcid.org/0009-0007-1449-3968

Косяков Сергей Витальевич, д.т.н., профессор, заведующий кафедрой программного обеспечения компьютерных систем, ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина» (153003, Россия, Иваново, ул. Рабфаковская, д. 34). E-mail: ksv@ispu.ru. Scopus Author ID 6507182528, ResearcherID H-5686-2018, SPIN-код РИНЦ 1371-9929, https://orcid.org/0000-0003-0231-0750

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt