УДК 004.934 DOI: 10.17587/it.25.313-318

Л. В. Савченко, канд. техн. наук, e-mail: lsavchenko@hse.ru,
Национальный исследовательский университет Высшая школа экономики — Нижний Новгород

Система постановки произношения на основе сверточных нейронных сетей и информационной теории восприятия речи

Рассматривается задача постановки произношения на основе применения методов глубокого обучения совместно с информационной теорией восприятия речи. Для повышения эффективности тестирования качества произношения предложено проводить дообучение сверточной нейронной сети с использованием наилучших эталонов пользователя. Экспериментально показано, что предложенный подход характеризуется высокой точностью и скоростью распознавания для различных акустических моделей по сравнению с известными аналогами.

***Ключевые слова**: постановка произношения, распознавание речи, сверточные нейронные сети, глубокое обучение, информационное рассогласование Кульбака — Лейблера*

Введение

Задача постановки произношения в последнее время вызывает интерес из-за обширной сферы практического применения [1, 2], например, в информационных системах обучения людей с нарушением слуха [3, 4] и особенностями речи (акцент, заикание), обучения иностранным языкам [5, 6] и в других областях, где существует необходимость постановки произношения, близкого к произношению некоторого эталонного диктора [7, 8]. Кроме того, качественное произношение слов является неотъемлемой частью задачи распознавания речи и способно существенно повысить точность распознавания [9].

Как известно, для различных носителей национального языка возникает проблема вариативности устной речи и тесно связанная с ней проблема самостоятельной оценки обучающимся качества своего произношения [10]. В самой постановке задачи содержится очевидное противоречие: обучаемый с недостаточной на данный момент языковой подготовкой и ограниченными возможностями в процессе самообучения должен приблизиться по своему произношению к некоторому эталону, который он слабо себе представляет. Указанное противоречие с успехом преодолевается с использованием предложенного в работе [11] подхода на основе информационной теории восприятия речи (ИТВР). В этом подходе достижимость эталонного произношения обеспечивается использованием лучших образцов произнесений от одного или даже группы дикторов. Система обучения на основе ИТВР способна запоминать лучшее произношение диктором слова и оценивать качество последующего произнесения того же слова по отношению к этим наилучшим для диктора словам, а не только по отношению к используемым по умолчанию эталонам, введенным идеальным диктором. Для оценки качества произношения используется тестирование различимости слов, которое может быть осуществлено с помощью одного из известных методов автоматического распознавания речи [12]. В указанных выше работах использовалась только реализация принципа минимума информационного рассогласования Кульбака—Лейблера [13] для сопоставления оценок спектральной плотности мощности речевых сигналов. В то же время сейчас, как известно [14], наибольшей точностью распознавания характеризуются подходы на основе технологий глубокого обучения и, в частности, сверточных нейронных сетей (СНС) [15]. Таким образом, представляет несомненный интерес комбинирование ИТВР и современных методов глубокого обучения для повышения эффективности систем постановки произношения. Исследованиям в этом актуальном направлении и посвящена настоящая работа.

Постановка произношения на основе информационной теории восприятия речи

Пусть задана база данных в виде множества из $R > 1$ эталонных слов $X_r = \{\mathbf{x}_{r,j}\}$, где $r = \overline{1,R}$ — номер слова, $\{\mathbf{x}_{r,j}\}$, $j = \overline{1,J_r}$ — вектор отсчетов речевого сигнала, представляющего j-ю реализацию r-го слова; J_r — число эталонных реализаций r-го слова. В традиционных системах обучения речи каждое слово, чаще всего, представляется одним словом ($J_r = 1$), произнесенным эталонным диктором [16]. Далее в процессе постановки произношения пользователь обычно последовательно обучается произносить каждое r-е слово до тех пор, пока вектор отсчетов $\mathbf{x}$ не будет достаточно близок к одному из эталонов:

$$\min_{j=\overline{1,J_r}} \rho(\mathbf{x}, \mathbf{x}_{r,j}) < \rho_0, \qquad (1)$$

где $\rho(\mathbf{x}, \mathbf{x}_{r,j})$ — некоторая мера близости между сигналами $\mathbf{x}$ и $\mathbf{x}_{r,j}$, а ρ_0 — подобранное экспериментально максимальное допустимое отклонение одноименных реализаций слов.

В подходе на основе ИТВР [11] достижимость эталонного произношения обеспечивается использованием не одного, а нескольких "эталонов", включающих в себя и лучшие образцы произнесений от одного или даже группы учащихся лиц, успешно прошедших обучение ранее. Такая система способна запоминать лучшее произношение диктором слова и оценивать качество последующего произнесения того же слова по отношению к этим наилучшим для диктора словам, а не только по отношению к используемым по умолчанию эталонам, введенным идеальным диктором.

Согласно ИТВР [11] одноименные реализации в сознании человека группируются в соответствующие классы вокруг некоторого центра — эталонной метки. Для определения понятия "центр кластера" в информационной теории восприятия речи используется кластерная модель слов [17]: эталон $\mathbf{x}_r^* \in X_r$ образует информационный центр r-го класса, если в пределах множества X_r он характеризуется минимальным разбросом — средней суммой информационных рассогласований Кульбака—Лейблера $\rho(\mathbf{x}_{r,k}, \mathbf{x}_{r,j})$ [18] относительно всех других его меток-реализаций $\mathbf{x}_{r,j}$, $j = \overline{1,J_r}$, в пределах каждого r-го класса:

$$\mathbf{x}_r^* = \arg\min_{\mathbf{x}_{r,k}, k \in \{1,\ldots,J_r\}} \delta_{r,k}, \qquad (2)$$

где $\delta_{r,k}$ — средний разброс эталонных реализаций r-го слова относительно k-го эталона:

$$\delta_{r,k} = \frac{1}{J_r} \sum_{j=1}^{J_r} \rho(\mathbf{x}_{r,k}, \mathbf{x}_{r,j}). \qquad (3)$$

После того как пользователь обучился произносить каждое слово, проводят оценку качества произношения всех слов с помощью их распознавания на основе принципа минимума информационного рассогласования [12]. Тем самым, каждое произнесенное слово будет характеризоваться не только близостью к эталону, но и его различимостью относительно других эталонов. В случае неудовлетворительного результата (плохая различимость определенных слов, их удаленность друг от друга) обучение произношению следует повторить.

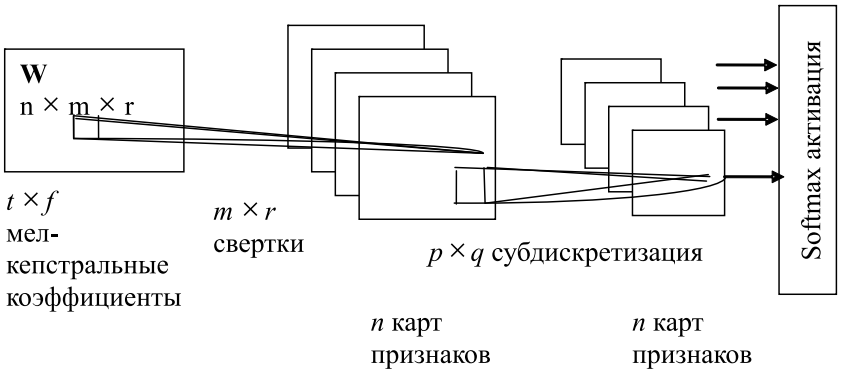
Заметим, что эффективность обучения во многом определяется качеством алгоритмов распознавания. В следующем разделе рассмотрим современные подходы к распознаванию изолированных слов.

Нейросетевые методы распознавания изолированных слов

Задача распознавания изолированных слов состоит в том, чтобы вновь поступающему на вход речевому сигналу поставить в соответствие одно из R эталонных слов. Одним из наиболее популярных способов решения задачи в настоящее время являются глубокие нейронные сети, которые предполагают моделирование акустических сигналов с помощью многоуровневых последовательно соединенных слоев нелинейных функций. Например, такие известные приложения голосового поиска, как Google и Apple Siri, реализованы с помощью глубоких сетей. Отметим, что при распознавании речи с помощью глубокой нейронной сети проводится множество матричных вычислений, что приводит к существенным затратам вычислительных ресурсов (хранение акустической модели) на этапе распознавания. При этом глубокие сети оказываются чувствительными к репрезентативности обучающего множества (речевого корпуса), поэтому их применение может привести к неудовлетворительным результатам для распознавания нестандартной речи (при наличии помех, дефектов произношения и т. п.) [19].

При работе с аудиосигналами в настоящее время, например в системах DeepSpeech [20],

вместо обычных многослойных нейронных сетей прямого распространения используются рекуррентные нейронные сети [21]. Такие сети имеют циклические связи и позволяют хранить информацию о предыдущем состоянии. Рекуррентные сети можно представить себе как множество копий одной и той же сети, причем каждая копия передает сообщение следующей копии. В задаче распознавания речи такие рекуррентные архитектуры позволяют автоматически настроить систему для эффективного различения набора распознаваемых слов [22].



Архитектура сверточной нейронной сети

К сожалению, вычислительная сложность распознавания речи с помощью глубоких рекуррентных сетей является чрезвычайно высокой, поэтому на практике для их реализаций обычно используются удаленные мощные вычислительные серверы. В связи с этим в последнее время приобретают популярность глубокие СНС, которые не только характеризуются намного более высокой скоростью принятия решений, но и показали более точные результаты распознавания на больших и маленьких словарях в ряде работ [22, 23]. Используемая в них свертка издавна применялась в цифровой обработке речевых сигналов в разнообразных методах линейной фильтрации. Особенность СНС заключается в том, что в ней нейроны первых уровней упорядочены в особую структуру — на первых слоях нейроны разбиты на карты признаков определенного размера, и разные карты внутри одного слоя соответствуют нейронам разного типа, которые реагируют на разные особенности спектра речевого сигнала.

Архитектура типичной СНС показана на рисунке. На вход сети подается речевой сигнал $\mathbf{V} \in \Re^{t \times f}$, где t и f — число фреймов и число характерных признаков. Например, в качестве входного сигнала могут использоваться мел-кепстральные коэффициенты размером $t \times f = 32 \times 40$. Затем трехмерный тензор весов $\mathbf{W} \in \Re^{(m \times r) \times n}$ сворачивается с входным сигналом **V**. Выход сверточного слоя затем передается на следующий сверточный слой. В библиотеке TensorFlow используется сверточная нейронная сеть типа cnn-trad-fpool3, которая, как показано в статье [24], позволяет повысить точность распознавания изолированных слов по сравнению с глубокими нейронными сетями на 27 %. На выходе СНС получаются оценки апостериорной вероятности принадлежности к каждому r-му эталонному классу слов с помощью softmax активации [25]:

$$P(r|\mathbf{x}) = \frac{\exp(z_r)}{\sum_{j=1}^{R} \exp(z_j)}, \quad r = 1, 2, \ldots, R. \qquad (4)$$

Здесь z_r — выход r-го нейрона предпоследнего слоя СНС, на вход которой подан сигнал **x**.

Предлагаемый в настоящей работе алгоритм постановки произношения на основе ИТВР и СНС представлен ниже.

Алгоритм постановки произношения на основе информационной теории восприятия речи и сверточных нейронных сетей

Входные данные: база данных эталонных слов $\mathbf{x}_r^*$.
Выходные данные: оценка точности и времени распознавания слов.
Параметры: порог на слова ρ_0, необходимое число добавляемых эталонов m.

1. Для каждого r-го класса, $r = \overline{1, R}$:
   1.1. Записать речевой сигнал **x**.
   1.2. Пока пользователь не обучится стабильному произношению, повторить (для каждого r-го класса)
      1.2.1. Для каждого произнесенного слова **x** с помощью предобученной СНС оценить апостериорную вероятность эталона r-го класса согласно (5).
   1.3. Пока число добавленных слов меньше m,
      1.3.1. Если максимальная апостериорная вероятность соответствует r-му классу, то
         1.3.1.1. Добавить слово **x** в базу данных эталонных слов.
2. Провести дообучение сверточной нейронной сети на основе добавленных слов.
3. Для каждого r-го класса слов повторить
   3.1. Определить информационный центр-эталон согласно (2).
   3.2. Записать речевой сигнал **x**.
   3.3. Оценить время, правильность распознавания слова **x** и средний разброс всех эталонов внутри класса согласно (3).
4. Вернуть среднюю вероятность правильного распознавания всех слов и время распознавания слова.

Здесь на первом этапе пользователь обучается стабильному произношению слов, т. е. в базу данных эталонов добавляются только те слова, для которых максимальная оценка апостериорной вероятности на выходе СНС (4) превышает некоторый наперед заданный порог p_0:

$$\max_{r=\overline{1,R}} P\left(r|\mathbf{x}\right) > p_0. \qquad (5)$$

Далее проводится дообучение СНС на основе наилучших пользовательских эталонов. После этого оценивается качество постановки произношения, т. е. вновь произнесенные слова распознаются с помощью дообученной СНС. Кроме того, для проверки стабильности произнесения оценивается степень близости произнесенных пользователем слов внутри r-го класса. Для этого в каждом r-м классе ищется центр-эталон (2) и вычисляется средний разброс ("радиус" в терминологии ИТВР) каждого класса — минимальная средняя удаленность от центра (3). Небольшой радиус соответствует качественному и стабильному произнесению слова. Если разброс оказывается больше некоторого наперед заданного порога, то обучение для r-го слова следует пройти еще раз.

Результаты экспериментальных исследований

В экспериментальной части статьи рассматривается задача оценки качества произношения десяти слов (команд) английского языка ("down", "go", "left", "no", "off", "on", "right", "stop", "up", "yes"). Дообучение СНС проводили с помощью набора скриптов Simple Audio Recognition из библиотеки TensorFlow. Использовали несколько встроенных акустических моделей. Модель conv базируется на топологии cnn-trad-fpool3 [24], модель low_latency_conv использует нейронную сеть cnn-one-fstride4 [24], модель low_latency_svdf использует топологию rank-constrained [26] (сжатие нейронных сетей), модель tiny_conv состоит из односверточного нейронного слоя (была разработана для работы на устройствах с небольшим объемом оперативной памяти). Экспериментально были оценены точность и время распознавания для всех доступных видов акустических моделей.

Для записи сигнала применялся встроенный в ноутбук микрофон. Частота дискретизации F установлена равной 16 кГц. Тестовое множество содержало 100 элементов (по 10 реализаций каждой команды). В табл. 1 представлены результаты сравнительного анализа точности и времени распознавания слов английского языка для всех доступных акустических моделей с обучением и без обучения (наилучшие результаты выделены полужирным шрифтом). Здесь в первой части эксперимента распознавание проводилось сначала для базовых акустических моделей на словах эталонного диктора (столбцы "Без дообучения" табл. 1). Во второй части эксперимента проводилось дообучение сверточной нейронной сети на 100 реализациях эталонных команд (по 10 реализаций каждой команды). В качестве эталонов были добавлены все 100 произносимых слов пользователя (столбцы "Все эталоны" в табл. 1). В заключение эксперимента дообучение сети проводилось на основе только лучших реализаций пользователя, выбранных с помощью критерия (5). Результаты приведены в столбцах "Лучшие эталоны" в табл. 1.

Из табл. 1 видно, что акустическая модель conv превосходит по точности распознавания другие доступные акустические модели, если не проводить обучение сети (72 %), в то же время наименьшее время распознавания команды имеет модель tiny_conv (4...5 мс). Дообучение

Таблица 1

Сравнительный анализ точности и времени распознавания для различных акустических моделей

Акустические модели	Точность и время распознавания					
	Без дообучения		С обучением			
			Все эталоны		Лучшие эталоны	
	Точность, %	Время, мс	Точность, %	Время, мс	Точность, %	Время, мс
Conv	**72**	7,5	91	6,5	94	6,5
low_latency_conv	27	7	**92**	6	**96**	6
low_latency_svdf	46	11	55	10.5	60	10.5
tiny_conv	20	**5**	50	**4**	65	**4**

Таблица 2

Разброс (3) эталонов для каждого слова

Слово	Радиус до обучения	Радиус после обучения
Down	1,9	0,17
Go	0,85	0,23
Left	0,55	0,3
No	1,14	0,25
Off	0,17	0,17
On	1,56	0,17
Right	0,54	0,3
Stop	0,28	0,28
Up	0,17	0,17

сети позволяет существенно повысить точность распознавания команд (на 19...65 %). В то же время обучение на лучших образцах пользователя позволяет повысить точность распознавания еще на 3...15 %. Таким образом, акустические модели conv и low_latency_conv позволяют достичь наивысшей точности распознавания, а акустическая модель tiny_conv имеет наименьшее время обработки речевого сигнала.

В качестве примера практического применения подхода к тестированию стабильности произнесения слов на заключительных шагах работы (см. *Алгоритм*) рассмотрим результаты оценки разбросов (3) внутри каждого класса (табл. 2). Первоначальная постановка произношения (второй столбец табл. 2) прошла недостаточно качественно (оценка радиуса (3) для слов "down", "no", "on" оказалась намного выше среднего значения). Поэтому, согласно предложенному подходу, пользователь повторно прошел обучение произнесению этих слов (табл. 2). Результаты (оценки среднего разброса (3)) представлены в третьем столбце табл. 2. Здесь качество постановки произношения всех слов можно считать приемлемым, так как разброс расстояний оказался значительно ниже по сравнению с первоначальными оценками радиусов (3). Таким образом, предлагаемый подход на основе ИТВР и глубоких СНС позволяет повысить точность распознавания слов за счет дообучения нейронной сети и постановки качественного произношения.

Заключение

В настоящей работе предложен новый подход для оценки качества произношения слов, основанный на комбинировании информационной теории восприятия речи и глубоких СНС, позволяющий увеличить точность распознавания слов за счет дообучения нейронной сети на наиболее качественно произнесенных словах пользователя. Результаты экспериментальных исследований показывают, что добавление в базу данных эталонных команд пользователя способно увеличить точность распознавания на 19...69 % (см. табл. 1). В дальнейших исследованиях интерес представляет применение предложенного подхода при распознавании команд при наличии у обучающегося пользователя явно выраженных дефектов речи (таких, например, как заикание, наличие акцента, отсутствие в речи некоторых звуков).

Список литературы

1. **Савченко В. В., Акатьев Д. Ю.** Обучение звуковому строю языка глухонемых и слабослышащих на основе информационной теории восприятия речи // Информационные технологии. 2010. № 2. С. 60—66.
2. **Krasnova E., Bulgakova E.** The use of speech technology in computer assisted language learning systems // Lecture Notes in Computer Science. 2014. Т. 8773. P. 459—466.
3. **Денисова И. А.** Игровые технологии как условие повышения качества произношения учащихся с нарушениями слуха младшего школьного возраста // Череповецкие научные чтения. 2015. С. 52—54.
4. **Schuller B., Steidlb S., Batlinerc A., Burkhardtd F., Devillersa L., Muller C., Narayananf S.** Paralinguistics in speech and language-State-of-the-art and the challenge // Computer Speech and Language. 2013. Vol. 27, N. 1. P. 4—39.
5. **Ахмедова М. М., Рахимова М. И., Отамуродова Ф. Э.** Методика обучения произношению иностранного языка // Наука и Мир. 2016. Т. 3, № 6 (34). С. 52—53.
6. **Golonka E.** Technologies for foreign language learning: a review of technology types and their effectiveness // Computer Assisted Language Learning. 2014. Т. 27, N. 1. P. 70—105.
7. **Miller C.** Computational Approaches to Exploring Persian-Accented English // Research in Language. 2015. Т. 13, N. 1. P. 51—60.
8. **Hu W., Qian Y., Soong F.** A new DNN-based high quality pronunciation evaluation for computer-aided language learning // INTERSPEECH. 2013. P. 1886—1890.
9. **Савченко Л. В.** Оценка качества произношения на основе метода нечеткого фонетического кодирования // Телекоммуникации. 2017. № 5. С. 42—48.
10. **Попова М. И.** Обучение студентов языкового вуза иноязычной письменной речи на основе информационных технологий // Наука и образование. 2011. № 4. С. 85—89.
11. **Савченко В. В.** Информационная теория обучения речи // Изв. вузов России. Радиоэлектроника. 2009. № 3. С. 3—12.
12. **Benesty J., Sondh M., Huang Y.** Springer handbook of speech recognition. New York: Springer, 2008, 1176 p.
13. **Kullback S.** Information theory and statistics. New York: Dover Pub, 1997. 408 p.
14. **LeCun Y., Bengio Y., Hinton G.** Deep learning // Nature. 2015. Т. 521, N. 7553. 436 p.
15. **Zhang Y., Chan W., Jaitly N.** Very deep convolutional networks for end-to-end speech recognition // Acoustics, Speech

and Signal Processing (ICASSP), 2017. IEEE International Conference on. IEEE, 2017. P. 4845—4849.

16. **Кнеллер Э. Г., Караульных Д. В.** Устройство для обучения разговорной (устной) речи с визуальной обратной связью: пат. на полезную модель № WO2016053141 A1, Роспатент: по заявке PCT/RU2015/000583 от 17.09.2015

17. **Савченко Л. В.** Автоматическое распознавание изолированных слов на основе теории нечетких множеств и кластерной модели минимальных речевых единиц // Информационные технологии. 2014. № 2. С. 9—13.

18. **Савченко В. В.** Исследование стационарности случайных временных рядов с использованием принципа минимума информационного рассогласования // Известия высших учебных заведений. Радиофизика. 2017. Т. 60, № 1. С. 89—96.

19. **Васильев Е. М., Меренков В. В.** Система распознавания фонетических образов на основе нейросетевой модели восприятия речи // Вестник Воронежского государственного технического университета. 2009. № 10. С. 130—134.

20. **Amodei D., Ananthanarayanan S., Anubhai R., Bai J., Battenberg E., Case C., Chen J.** Deep speech 2: End-to-end speech recognition in english and mandarin // International Conference on Machine Learning. 2016. С. 173—182.

21. **Graves A., Mohamed A., Hinton G.** Speech recognition with deep recurrent neural networks //Acoustics, speech and signal processing (ICASSP). 2013. IEEE, 2013. P. 6645—6649.

22. **Toth L.** Combining Time-and Frequency-Domain Convolution in Convolutional Neural Network-Based Phone Recognition // Proceedings of the Acoustics, speech and signal processing (ICASSP). 2014. P. 190—194.

23. **Sainath T. N., Mohamed A., Kingsbury B., Ramabhadran B.** Deep Convolutional Neural Networks for LVCSR // Proceedings of the Acoustics, speech and signal processing (ICASSP). 2013. P. 8614—8618.

24. **Sainath T. N., Parada C.** Convolutional neural networks for small-footprint keyword spotting // Sixteenth Annual Conference of the International Speech Communication Association (ICASSP). 2015. P. 1478—1482.

25. **Zhang Y., Pezeshki M., Brakel P., Zhang S., Bengio C., Courville A.** Towards end-to-end speech recognition with deep convolutional neural networks //arXiv preprint arXiv:1701.02720, 2017.

26. **Nakkiran P., Alvarez R., Prabhavalkar R., Parada C.** Compressing deep neural networks using a rank-constrained topology // Sixteenth Annual Conference of the International Speech Communication Association (ICASSP). 2015. P. 1473—1477.

L. V. Savchenko, Ph. D., e-mail: lsavchenko@hse.ru,
National Research University Higher School of Economics — N. Novgorod

Computer-Assisted Language Learning Based on Convolutional Neural Networks and Information Theory of Speech Perception

In this paper we consider a problem of computer assisted language and pronunciation learning based on the deep neural networks and the information theory of speech perception. At first, a user learns the stable pronunciation of words. The best utterances from the user with high posterior probability estimated by the pre-trained convolutional neural network are added to the training set. Next, this training set is used to fine-tune this convolutional neural network. If new utterances are successfully recognized with the resulted neural network, it is concluded that pronunciation of all words is distinguishable. In this case in order to additionally verify the stability of pronunciation of each class (word), the closeness of the user pronunciations is estimated by computing the average Kullback-Leibler information discrimination between each signal and the centroid reference of the class. If this mean discrimination for particular word is greater than a certain threshold, then the training for this word should be repeated. The experimental results for learning of English words proved that the proposed approach is characterized by higher accuracy and speed for existing acoustic models when compared to conventional techniques.

Keywords: *computer-assisted learning system, speech recognition, convolutional neural network, deep learning, Kullback-Leibler information discrimination*

DOI: 10.17587/it.25.313-318

References

1. **Savchenko V. V., Akatjev D. Yu.** Learning the sound structure of the language of deaf and hard of hearing on the basis of information theory of speech perception, *Informacionnye Tekhnologii*, 2010, no. 2, pp. 60—66 (in Russian).

2. **Krasnova E., Bulgakova E.** The use of speech technology in computer assisted language learning systems, *Lecture Notes in Computer Science*, 2014, vol. 8773, pp. 459—466.

3. **Denisova I. A.** Game technologies as a condition of improving the quality of pronunciation of students with hearing impairments of primary school age, *Cherepovets Scientific Readings*, 2015, pp. 52—54 (in Russian).

4. **Schuller B., Steidlb S., Batlinerc A., Burkhardtd F., Devillersa L., Muller C., Narayananf S.** Paralinguistics in speech and language-State-of-the-art and the challenge, *Computer Speech and Language*, 2013, vol. 27, no. 1, pp. 4—39.

5. **Ahmedova M. M., Rahimova M. I., Otamurodova F.Je.** Methods of teaching foreign language pronunciation, *Science and World*, 2016, vol. 3, no. 6 (34), pp. 52—53 (In Russian).

6. **Golonka E.** Technologies for foreign language learning: a review of technology types and their effectiveness, *Computer Assisted Language Learning*, 2014, vol. 27, no. 1, pp. 70—105.

7. **Miller C.** Computational Approaches to Exploring Persian-Accented English, *Research in Language*, 2015, vol. 13, no. 1, pp. 51—60.

8. **Hu W., Qian Y., Soong F.** A new DNN-based high quality pronunciation evaluation for computer-aided language learning, *INTERSPEECH*, 2013, pp. 1886—1890.

9. **Savchenko V. V.** Pronunciation quality assessment based on the fuzzy phonetic coding method, *Telecommunications and Radio Engineering*, 2017, no. 5, pp. 42—48 (in Russian).

10. **Popova M. I.** Learning students of the language University of foreign language writing on the basis of information technology, *Science and Education*, 2011, no. 4, pp. 85—89 (in Russian).

11. **Savchenko V. V.** Information theory of speech learning, *Journal of the Russian Universities. Radioelectronics*, 2009, no. 3, pp. 3—12 (in Russian).

12. **Benesty J., Sondh M., Huang Y.** Springer handbook of speech recognition, New York, Springer, 2008, 1176 p.

13. **Kullback S.** Information theory and statistics, New York, Dover Pub, 1997, 408 p.

14. **LeCun Y., Bengio Y., Hinton G.** Deep learning, *Nature,* 2015, vol. 521, no. 7553, 436 p.

15. **Zhang Y., Chan W., Jaitly N.** Very deep convolutional networks for end-to-end speech recognition, *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, 2017, pp. 4845—4849.

16. **Kneller E. G., Karaulnykh D. V.** A device for learning conversational (oral) speech with visual feedback: Patent RF № WO2016053141 A1, Rospatent: by request PCT/RU2015/000583 от 17.09.2015 (in Russian).

17. **Savchenko L. V.** Automatic recognition of isolated words on the basis of fuzzy set theory and cluster model of minimal speech units, *Informacionnye Tekhnologii*, 2014, no.2, pp. 9—13 (in Russian).

18. **Savchenko V. V.** Study of the stationarity of random time series using the principle of the information-divergence minimum, *Radiophysics and Quantum Electronics*, 2017, vol. 60, no. 1, pp. 81—87.

19. **Vasilyev E. M., Merenkov V. V.** Vowels recognition system on the basis of neuronet simulation of speech perception, *Herald of Voronezh state technical University*, 2009, no. 10, pp. 130—134 (in Russian).

20. **Amodei D., Ananthanarayanan S., Anubhai R., Bai J., Battenberg E., Case C., Chen J.** Deep speech 2: End-to-end speech recognition in English and Mandarin, International Conference on Machine Learning, 2016, pp. 173—182.

21. **Graves A., Mohamed A., Hinton G.** Speech recognition with deep recurrent neural networks, *Acoustics, speech and signal processing (ICASSP)*, 2013, IEEE.— 2013, pp. 6645—6649.

22. **Toth L.** Combining Time-and Frequency-Domain Convolution in Convolutional Neural Network-Based Phone Recognition, *Proceedings of the Acoustics, speech and signal processing (ICASSP),* 2014, pp. 190—194.

23. **Sainath T. N., Mohamed A., Kingsbury B., Ramabhadran B.** Deep Convolutional Neural Networks for LVCSR, *Proceedings of the Acoustics, speech and signal processing (ICASSP),* 2013, pp. 8614—8618.

24. **Sainath T. N., Parada C.** Convolutional neural networks for small-footprint keyword spotting, *Sixteenth Annual Conference of the International Speech Communication Association (ICASSP),* 2015, pp. 1478—1482.

25. **Zhang Y., Pezeshki M., Brakel P., Zhang S., Bengio C., Courville A.** Towards end-to-end speech recognition with deep convolutional neural networks, arXiv preprint arXiv:1701.02720, 2017.

26. **Nakkiran P., Alvarez R., Prabhavalkar R., Parada C.** Compressing deep neural networks using a rank-constrained topology, *Sixteenth Annual Conference of the International Speech Communication Association (ICASSP),* 2015, pp. 1473—1477.

Адрес редакции:

107076, Москва, Стромынский пер., 4

Телефон редакции журнала **(499) 269-5510**
E-mail: it@novtex.ru

Технический редактор *Е. В. Конова.*
Корректор *Е. В. Комиссарова.*
Сдано в набор 12.03.2019. Подписано в печать 24.04.2019. Формат 60×88 1/8. Бумага офсетная.
Усл. печ. л. 8,86. Заказ IT519. Цена договорная.

Журнал зарегистрирован в Министерстве Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций.
Свидетельство о регистрации ПИ № 77-15565 от 02 июня 2003 г.

Оригинал-макет ООО "Адвансед солюшнз". Отпечатано в ООО "Адвансед солюшнз".
119071, г. Москва, Ленинский пр-т, д. 19, стр. 1.