УДК 004.934 DOI: 10.17587/it.26.290-296

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

Распознавание изолированных слов на основе взвешенного голосования дикторозависимых нейросетевых моделей

Рассматривается задача распознавания изолированных слов с помощью методов глубокого обучения и сверточных нейронных сетей. Предложено выполнить дообучение сетей для проведения адаптации акустических моделей на голос диктора с использованием малого числа произнесенных им реализаций эталонных слов. Для понижения вероятности ошибочного распознавания рассматривается комбинирование нескольких различных дообученных дикторозависимых нейросетевых моделей.

***Ключевые слова**: распознавание речи, сверточные нейронные сети, глубокое обучение, ансамбль моделей, адаптация акустической модели, взвешенное голосование*

Введение

Задача распознавания речи в последнее время вызывает интерес из-за обширной сферы практического применения [1—3], например, в информационных системах поиска информации, диктовки текста, интерактивного речевого взаимодействия, при обучении людей с нарушением слуха [4, 5] и особенностями речи (акцент, заикание). Однако применение существующих систем распознавания на практике ограничено недостаточно высокой степенью надежности их функционирования в условиях интенсивных акустических помех, таких как шум улицы, звуки от проезжающего транспорта, речь посторонних лиц. Даже при малой интенсивности такого рода помех эффективность систем распознавания, как правило, существенно снижается [6]. В настоящее время наибольшей точностью распознавания характеризуются способы формирования акустических моделей на основе технологий глубокого обучения и, в частности, сверточных нейронных сетей (СНС) [7]. Известно большое число различных архитектур СНС, применимых для распознавания речи, точность которых может существенно варьироваться в зависимости от уровня помех и особенностей речи конкретного диктора.

Для задач обработки изображений хорошо изучена возможность адаптации СНС к новой предметной области за счет дообучения с использованием обучающих выборок малого объема [8, 9]. Поэтому для повышения надежности распознавания изолированных слов и фраз в настоящей статье предлагается выполнить дообучение СНС на голос нового пользователя, что особенно важно для дикторов с дефектами речи (акцент, заикание, отсутствие в речи некоторых звуков) и при распознавании неродной речи [10]. Для дальнейшего повышения точности предлагается рассматривать формирование ансамблей таких дообученных акустических моделей [11, 12]. Полученные результаты и сделанные по ним выводы представляют интерес для широкого круга специалистов в области автоматического распознавания речи.

Нейросетевые методы распознавания изолированных слов

Задача распознавания слов состоит в том, чтобы входному (распознаваемому) слову X поставить в соответствие наиболее близкое слово из словаря, содержащего R различных слов [1]. Одним из наиболее популярных методов решения задачи в настоящее время является использование глубоких нейронных сетей [13], 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 рекуррентных сетей. Отметим, что при этом выполняется множество матричных вычислений, а это приводит к существенным затратам вычислительных ресурсов (хранение акустической модели) на этапе распознавания. 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.) [1].

Поэтому в последнее время все чаще применяются 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 [14, 15]. 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. Например, в библиотеке TensorFlow используется СНС типа cnn-trad-fpool3, которая позволяет повысить точность распознавания изолированных слов по сравнению с глубокими нейронными сетями на 27 % [16]. На выходе СНС получаются оценки апостериорной вероятности принадлежности к каждому r-му слову с помощью softmax активации [17]:

$$P(r|X) = \frac{\exp(z_r)}{\sum\limits_{j=1}^{R} \exp(z_j)}, \quad r = 1, 2, \ldots, R. \qquad (1)$$

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

Большинство современных систем распознавания речи ориентируются на дикторонезависимое распознавание речи в благоприятных условиях, в которых акустические помехи сведены к минимуму [6]. В то же время результаты распознавания могут значительно ухудшаться, например, при наличии во входном сигнале помех или при существенных отличиях голоса пользователя от большинства голосов дикторов в речевой базе данных, использующейся для обучения СНС. Поиску путей повышения точности нейросетевых моделей и посвящена настоящая статья.

Предложенный подход

Прежде всего отметим, что вычислительная сложность нейросетевых моделей, применяющихся для распознавания изолированных слов [16], является достаточно низкой. Поэтому на практике для повышения точности распознавания можно использовать традиционные для распознавания образов комитеты классификаторов на основе ансамбля нескольких различных моделей [12, 18]. Рассмотрим далее наиболее часто применяющиеся ансамбли, не требующие наличия дополнительной большой обучающей выборки. Пусть на вход K акустических моделей подается слово X, и в результате на выходе каждой из них получаются оценки апостериорной вероятности $P_k(r|X)$, $k = \overline{1, K}$, $r = \overline{1, R}$ (1). Тогда решение принимается в пользу r-го слова согласно одному из критериев:

1. *Максимум апостериорной вероятности*

$$r^* = \arg\max_{r=\overline{1,R}} \max_{k=\overline{1,K}} P_k(r|X). \qquad (2)$$

2. *Максимум средней апостериорной вероятности* (sum rule [18])

$$r^* = \arg\max_{r=\overline{1,R}} \frac{1}{K} \sum_{k=1}^{K} P_k(r|X). \qquad (3)$$

3. Если предположить, что все K моделей независимы, то можно оценить итоговую апостериорную вероятность как произведение выходов (1) каждой модели. Такое правило (product rule [18]) в вычислительном плане обычно записывается с использованием *логарифма средней апостериорной вероятности:*

$$r^* = \arg\max_{r=\overline{1,R}} \log \frac{1}{K} \sum_{k=1}^{K} P_k(r|X). \qquad (4)$$

4. В работах [19, 20] показаны преимущества построения ансамблей на основе принципа *минимума информационного рассогласования (МИР) Кульбака—Лейблера*

$$k^* = \arg\min_{k=\overline{1,K}} \sum_{j=1}^{K} \sum_{r=1}^{R} \left(P_k(r|X) \log \frac{P_k(r|X)}{P_j(r|X)} \right). \qquad (5)$$

Решение принимается с помощью k^*-модели:

$$r^* = \arg\max_{r=\overline{1,R}} P_{k^*}(r|X).$$

В описанных выше подходах предполагается, что все члены ансамбля вносят одинаковый вклад в итоговое решение. В то же время, как известно, точность распознавания с помощью различных СНС может существенно различаться. Поэтому в настоящей работе использу-

ются методы взвешенного голосования [19], в которых вес w_k, $k = \overline{1,K}$, каждой k-й модели предлагается определить пропорционально оценке ее точности на контрольном множестве при введении нормировки $\sum_{k=1}^{K} w_k = 1$. Чем точнее модель, тем больший вклад она вносит в итоговое решение. Будем рассматривать метод *максимума взвешенной суммы апостериорной вероятности*

$$r^* = \arg\max_{r=\overline{1,R}} \sum_{k=1}^{K} (w_k P_k(r|X)), \qquad (6)$$

а также, по аналогии с (4), в предположении о независимости акустических моделей, критерий *максимума логарифма взвешенной суммы апостериорной вероятности*

$$r^* = \arg\max_{r=\overline{1,R}} \log \sum_{k=1}^{K} (w_k P_k(r|X)). \qquad (7)$$

К сожалению, даже применение ансамбля акустических моделей не приведет к значимому повышению надежности распознавания в случае, если входной сигнал существенно отличается от образцов в обучающей выборке, например, при наличии акустических помех или для дикторов с выраженными особенностями произношения, в том числе для неродного языка. Поэтому в данной статье предлагается по аналогии с доменной адаптацией для задач распознавания изображений [8, 9] выполнить настройку акустических моделей в ансамбле на голос конкретного пользователя. Для этого требуется сформировать небольшую выборку из 1...5 образцов произношений каждого из R слов.

На рис. 1 представлен предложенный алгоритм распознавания слов с адаптацией акустических моделей на голос диктора с использованием малого числа произнесенных им реализаций эталонных слов и комбинированием нескольких различных дообученных таких способом дикторозависимых нейросетевых моделей. Здесь для отбраковки некачественных речевых сигналов предлагается предварительно отбирать такие образцы, которые правильно распознаются исходной (дикторонезависимой) СНС. Далее стандартными средствами выполняется дообучение K акустических моделей с использованием собранного обучающего множества [8], после чего для принятия итогового решения применяется взвешенное голосование (5) или (6).

Входные данные: распознаваемое слово X, валидационное множество всех слов, K СНС (дикторонезависимых акустических моделей)
Выходные данные: метка класса слова из словаря 1,2,…R
Параметры: число добавляемых эталонов m каждого слова
1. Для каждого r-го слова ($r = \overline{1,R}$) повторить
  1.1. Пока число добавленных слов меньше m
    1.1.1. Записать речевой сигнал X_r.
    1.1.2. Подать X_r на вход наилучшей СНС и оценить апостериорные вероятности (1).
    1.1.3. Если максимальная апостериорная вероятность соответствует r-му классу, то
      1.1.3.1. Добавить слово X_r в базу данных эталонных слов.
3. Для $k = \overline{1,K}$ повторить
  3.1. Провести дообучение k-й акустической модели (СНС) на основе малого числа m добавленных слов.
  3.2. Оценить с помощью дообученной СНС точность P_k распознавания валидационного множества слов.
4. Для $k = \overline{1,K}$ повторить
  4.1. Вычислить вес акустической модели $w_k = \frac{P_k}{\sum_{k=1}^{K} P_k}$.
5. Для каждого распознаваемого слова X
  5.1. Для каждой k-й акустической модели вычислить апостериорные вероятности $P_k(r|X)$.
6. Вернуть ближайшее эталонное слово r^* с помощью взвешенного голосования (5) или (6).

Рис. 1. Предлагаемый алгоритм распознавания слов на основе взвешенного голосования дикторозависимых нейросетевых моделей

Результаты экспериментальных исследований

В экспериментальной части статьи рассматривается задача оценки точности распознавания R = 10 слов английского языка ("zero", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine") в зависимости от уровня шума. Дообучение СНС проводили с помощью набора сценариев Simple Audio Recognition из библиотеки TensorFlow. В эксперименте использовалось несколько встроенных акустических моделей. Модель conv базируется на топологии cnn-trad-fpool3 [16], модель low_latency_conv использует нейронную сеть cnn-one-fstride4 [16], модель low_latency_svdf использует топологию rank-constrained [21] (сжатие нейронных сетей) и модель tiny_conv состоит из односверточного нейронного слоя (была разработана для работы на устройствах с небольшим объемом оперативной памяти) [16].

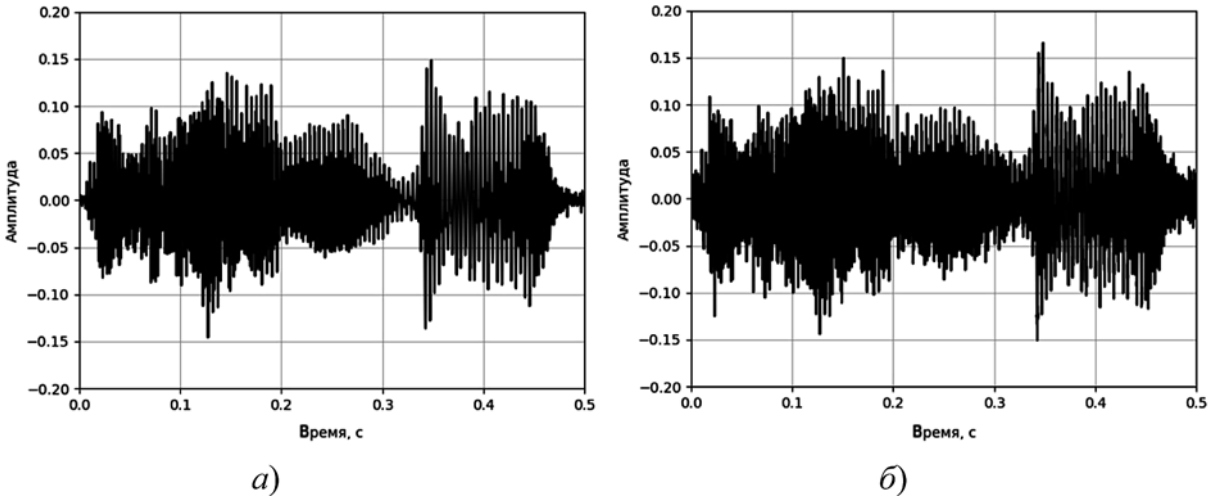
Для записи речевых сигналов применяли встроенный в ноутбук микрофон. Частота дискретизации F была установлена равной 16 кГц. В соответствии с предложенным подходом (рис. 1) формировалось обучающее множество из 10 слов (по одной реализации каждого слова) Для каждой из четырех СНС веса w_k определялись в соответствии с оценкой точностью модели, полученной для обучающего множества. Были получены следующие значения весов: для модели conv $w_1 = 0{,}3$, для low_latency_conv $w_2 = 0{,}3$, для low_latency_svdf $w_3 = 0{,}25$ и для tiny_conv $w_4 = 0{,}15$.

Тестовое множество содержало 200 реализаций речевых сигналов (по 20 реализаций каждого слова). К записанным в идеальных условиях речевым сигналам добавлялся аддитивный гауссовский шум различной амплитуды. На рис. 2 представлена временная диаграмма речевого сигнала слова "down" с разным уровнем аддитивного гауссовского шума, а на рис. 3 — оценки спектральной плотности мощности этих сигналов.

Из рис. 2, 3 видно, что наличие сильного шума приводит к изменениям спектральных характеристик, что очевидным образом сказывается на результатах распознавания.

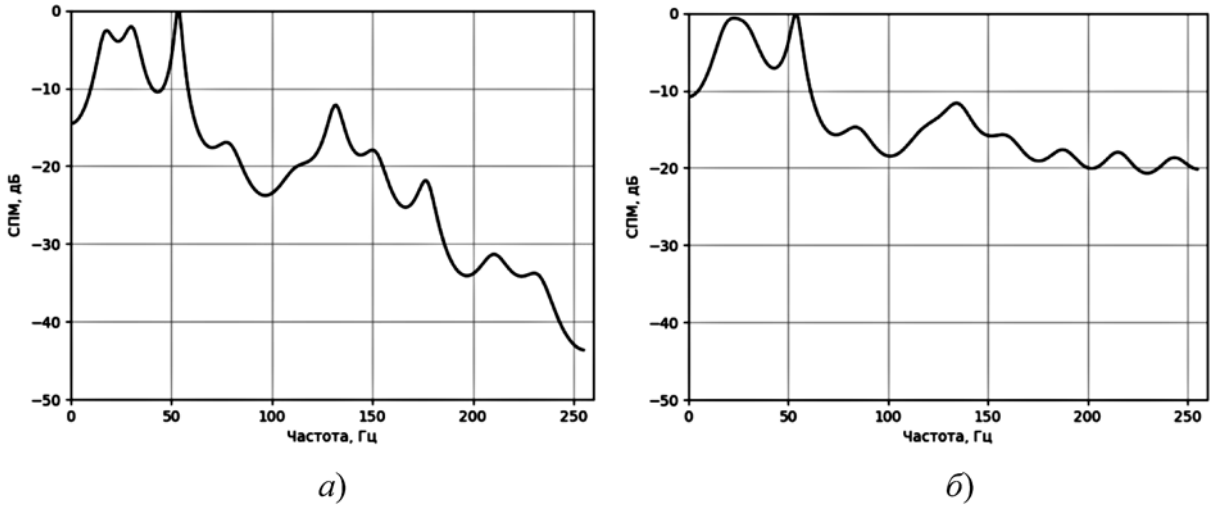
В табл. 1 представлены результаты сравнительного анализа точности распознавания слов для всех доступных традиционных акустических моделей и их ансамблей без дообучения СНС и с дообучением СНС. Наилучшие результаты выделены полужирным шрифтом.

Из табл. 1 видно, что предложенное дообучение СНС даже на небольшом числе образцов (всего 10 эталонных слов) способно повысить



а) *б*)

Рис. 2. Временная диаграмма речевого сигнала для слова "down":
а — отношение сигнал/шум 30 дБ; *б* — отношение сигнал/шум 10 дБ



а) *б*)

Рис. 3. Оценки спектральной плотности мощности (СПМ) речевого сигнала "down":
а — отношение сигнал/шум 30 дБ; *б* — отношение сигнал/шум 10 дБ

Таблица 1

Сравнительный анализ точности распознавания (%) для традиционных акустических моделей без дообучения СНС и с дообучением СНС

Акустические модели	Отношение сигнал/шум, дБ						
	30	25	20	15	10	5	0
Conv	65,16	65,33	65,99	60,79	57,80	48,99	39,56
Low_conv	52,24	48,41	46,25	45,59	41,58	38,44	36,63
Low_svdf	66,71	63,19	65,29	58,03	55,43	40,88	40,69
Tiny	43,80	42,39	43,43	39,04	35,32	30,32	21,40
Conv с обучением	67,02	**67,02**	**69,97**	**63,19**	61,80	**56,60**	**47,98**
Low_conv с обучением	53,58	50,83	49,98	49,26	47,76	43,93	41,78
Low_svdf с обучением	**67,64**	65,02	67,19	62,83	**62,87**	49,88	45,69
Tiny с обучением	56,58	54,83	53,33	52,34	50,59	48,20	39,11

Таблица 2

Сравнительный анализ точности распознавания (%) для для ансамбля акустических моделей без дообучения СНС и с дообучением СНС

Акустические модели	Отношение сигнал/шум, дБ						
	30	25	20	15	10	5	0
Максимум вероятности (2)	68,54	69,05	67,23	65,17	61,24	51,65	38,49
Sum rule (3)	67,45	68,03	67,14	66,54	62,12	53,14	42,23
Product rule (4)	69,11	69,02	68,41	66,78	61,91	52,34	41,05
МИР (5)	67,98	68,02	67,37	67,13	63,69	52,82	40,23
Предложенный подход (6)	71,23	72,01	70,32	69,16	65,17	55,34	42,12
Предложенный подход (7)	70,83	71,09	70,12	69,62	64,12	54,47	42,08
Максимум вероятности (2) с обучением	69,34	70,05	68,76	67,29	63,45	48,14	42,56
Sum rule (3) с обучением	72,23	69,34	68,78	68,43	64,32	54,23	48,89
Product rule (4) с обучением	72,67	70,23	68,69	68,14	64,35	54,78	49,12
МИР (5) с обучением	70,23	70,12	68,75	68,54	68,08	55,43	49,15
Предложенный подход (6) с обучением	**74,94**	**74,52**	**72,66**	71,24	66,77	**59,88**	**52,29**
Предложенный подход (7) с обучением	73,43	72,38	71,35	**72,15**	**68,26**	57,19	50,44

точность распознавания на 2...8 %, что особенно заметно при наличии сильных помех. Так при отношении сигнал/шум 0 дБ точность возросла на 6...18 %.

Из табл. 2 можно заметить, что точность распознавания можно повысить на 5...7 % по сравнению с традиционными акустическими моделями (табл. 1), если в итоговом решении в задаче распознавания слов применять ансамбль акустических моделей на основе взвешенного голосования.

Заключение

В настоящей работе для повышения точности распознавания слов предложено, во-первых, проводить адаптацию акустических моделей на голос диктора с использованием малого числа произнесенных им реализаций эталонных слов. Во-вторых, для понижения вероятности ошибочного распознавания рассматривается комбинирование в одном ансамбле нескольких различных дообученных дикторо-

зависимых нейросетевых моделей. Итоговое решение в задаче распознавания принимается с помощью взвешенного голосования, при этом для определения весов используются оценки точности каждой акустической модели для обучающей выборки. Предложенный подход (см. рис. 1) позволяет существенно повысить точность распознавания, особенно при наличии акустических помех. Проведенные экспериментальные исследования показывают, что использование ансамбля дообученных акустических моделей (табл. 2) позволяет повысить точность распознавания по сравнению с традиционными методами на 5...7 %. 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. **Benesty J., Sondh M., Huang Y.** Springer handbook of speech recognition. New York: Springer, 2008. 1176 p.
2. **Савченко Л. В.** Система постановки произношения на основе сверточных нейронных сетей и информационной теории восприятия речи // Информационные технологии. 2019. Т. 25. № 5. С. 313—318.
3. **Савченко А. В., Савченко В. В.** Метод измерения частоты основного тона речевого сигнала для систем акустического анализа речи // Измерительная техника. 2019. № . 3. С. 59—63.
4. **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.
5. **Денисова И. А.** Игровые технологии как условие повышения качества произношения учащихся с нарушениями слуха младшего школьного возраста // Череповецкие научные чтения. 2015. С. 52—54.
6. Савченко В. В. Распознавание речи на фоне шума методом фонетического декодирования слов // Телекоммуникации. 2016. № 9. С. 9—16.
7. **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.
8. Goodfellow I., Bengio Y., Courville A. Deep learning // MIT press. 2016. P. 781.
9. **Savchenko A.** Sequential three-way decisions in multi-category image recognition with deep features based on distance factor // Information Sciences. 2019. Vol. 489. P. 18—36.
10. **Tao J., Ghaffarzadegan S., Chen L., Zechner K.** Exploring deep learning architectures for automatically grading non-native spontaneous speech // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. P. 6140—6144.
11. Tresp V. Committee machines, Handbook for Neural Network Signal Processing. CRC Press. 2001. P. 135—151.
12. Savchenko A. V. Adaptive Video Image Recognition System Using a Committee Machine // Optical Memory and Neural Networks (Information Optics). 2012. Vol. 21, N. 4. P. 219—226.
13. **Hinton G. et al.** Deep neural networks for acoustic modeling in speech recognition //IEEE Signal processing magazine. 2012. T. 29. P. 82—97.
14. **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.
15. **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.
16. **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.
17. Liu W., Wen Y., Yu Z., Yang M. Large-margin softmax loss for convolutional neural networks // Proceedings of the International Conference on Machine Learning (ICML). 2016. Vol. 2, N. 3. P. 7.
18. **Kittler J., Hatef M., Duin R. Matas J.** On combining classifiers // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. V. 20, N. 3. P. 226—234.
19. **Theodoridis S., Konstantinos K.** Pattern recognition. Elsevier, 2008.
20. **Савченко А. В.** Образ как совокупность выборок независимых одинаково распределенных значений признаков в задачах распознавания сложноструктурированных объектов // Заводская лаборатория. Диагностика материалов. 2014. Т. 80, № 3. С. 70.
21. **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, PhD (Candidate of Sciences), e-mail: lsavchenko@hse.ru,
National Research University Higher School of Economics, N. Novgorod, Russian Federation

Isolated Words Recognition Based on Weighted Voting of Speaker-Dependent Neural Network Acoustic Models

The article deals with the problem of isolated words recognition based on deep convolutional neural networks. The use of existing recognition systems in practice is limited by an insufficiently high degree of their reliability functioning in conditions of intense acoustic noise, such as street noise, sounds from passing vehicles, etc. Nowadays, the most accurate recognition methods are characterized by the formation of acoustic models with deep learning technologies and, in particular, convolutional neural networks. For image processing problems the possibility of adaptation of such networks to a new domain with additional fine-tuning on rather small training samples is well studied. In this paper we proposed to perform additional training of networks

for adaptation of acoustic models on a speaker voice with use of small number of the utterances. In order to reduce the error rate, we consider an ensemble of several different speaker-dependent neural network architectures that have been trained in such a way. The final decision is made by a weighted voting rule, in which the weight of each acoustic model is determined in proportion to the accuracy estimated on the training set. The experimental results for recognition of English commands proved that such ensemble of pre-trained acoustic models can significantly improve accuracy compared to traditional pre-trained models, especially if the white Gaussian noise is added to the input signal.

Keywords: *speech recognition, isolated words recognition, convolutional neural networks, deep learning, ensemble of neural networks, acoustic model adaptation, weighted voting*

DOI: 10.17587/it.26.290-296

References

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

2. **Savchenko L. V.** Computer-assisted language learning based on convolutional neural networks and information theory of speech perception, *Information Technologies,* 2019, vol. 25, no. 5, pp. 313—318 (in Russian).

3. **Savchenko A. V., Savchenko V. V.** A Method for Measuring the Pitch Frequency of Speech Signals for the Systems of Acoustic Speech Analysis, *Measurement Techniques,* 2019, vol. 62, no. 3, pp. 282—288.

4. **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, *Information Technologies*, 2010, no. 2, pp. 60—66 (in Russian).

5. **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).

6. Savchenko V. V. Speech recognition in background noise by the method of the phonetic words decoding, *Telecommunications and Radio Engineering*, 2016, no. 9, pp. 9—16.

7. **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.

8. Goodfellow I., Bengio Y., Courville A. Deep learning, MIT press, 2016, p. 781.

9. **Savchenko A. V.** Sequential three-way decisions in multi-category image recognition with deep features based on distance factor, *Information Sciences*, 2019, vol. 489, pp. 18—36.

10. **Tao J., Ghaffarzadegan S., Chen L., Zechner K.** Exploring deep learning architectures for automatically grading non-native spontaneous speech, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2016, pp. 6140—6144.

11. **Tresp V.** Committee machines, Handbook for Neural Network Signal Processing, CRC Press, 2001, pp. 135—151.

12. **Savchenko A. V.** Adaptive Video Image Recognition System Using a Committee Machine, Optical Memory and Neural Networks (Information Optics), 2012, vol. 21, no. 4, pp. 219—226.

13. **Hinton G.** et al. Deep neural networks for acoustic modeling in speech recognition, *IEEE Signal processing magazine*, 2012, vol. 29, pp. 82—97.

14. **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.

15. **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.

16. **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.

17. **Liu W., Wen Y., Yu Z., Yang M.** Large-margin softmax loss for convolutional neural networks, *ICML,* 2016, vol. 2, no. 3, p. 7.

18. **Kittler J., Hatef M., Duin R., Matas J.** On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, vol. 20, no. 3, pp. 226—234.

19. **Theodoridis S., Konstantinos K.** Pattern recognition, Elsevier, 2008.

20. **Savchenko A. V.** The image as a set of samples of independent equally distributed values of features in the problems of recognition of complex structured objects, *Zavodskaya Laboratoriya*, 2014, vol. 80, no. 3, pp. 70 (in Russian).

21. **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.