НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2018 Том 18 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2018 Vol. 18 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.93

ОБЪЕДИНЕНИЕ ПРИЗНАКОВ В ЗАДАЧЕ ОБУЧЕНИЯ НЕЙРОСЕТЕВЫХ АКУСТИЧЕСКИХ МОДЕЛЕЙ

А.Н. Романенко[a,b,c]

[a] ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация
[b] Ульмский университет, Ульм, 89081, Германия
[c] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: anromanenko@corp.ifmo.ru

Информация о статье
Поступила в редакцию 17.01.18, принята к печати 20.02.18
doi: 10.17586/2226-1494-2018-18-2-350-352
Язык статьи – русский

Ссылка для цитирования: Романенко А.Н. Объединение признаков в задаче обучения нейросетевых акустических моделей // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 350–352. doi: 10.17586/2226-1494-2018-18-2-350-352

Аннотация
Предложен метод объединения признаков для задачи обучения нейросетевых акустических моделей с целью повышения качества распознавания речи. В отличие от способа подачи на вход нейронной сети конкатенированного вектора признаков различной природы, предлагаемый метод использует отложенное объединение на уровне скрытых слоев. Оно реализуется за счет использования индивидуальных входных потоков для каждого типа признаков. Такие потоки способны извлекать паттерны, характерные для каждого типа признаков, а затем объединять их на скрытом слое нейросетевой акустической модели. Влияние метода на качество системы было исследовано в задаче распознавания телефонной русской речи. Предложенный метод позволил добиться 0,41% абсолютного уменьшения пословной ошибки распознавания относительно конкатенации признаков и 1,35% в сравнении с наилучшей системой, использующей один вид признаков. Результаты работы могут быть использованы при разработке систем автоматического распознавания речи.

Ключевые слова
объединение признаков, нейросетевые акустические модели, распознавание речи

Благодарности
Работа выполнена при поддержке Министерства образования и науки Российской Федерации, госзадание № 8.9971.2017/ДААД.

FEATURE COMBINATION FOR THE TASK OF NEURAL NETWORK ACOUSTIC MODEL LEARNING

A.N. Romanenko[a,b,c]

[a] STC Ltd., Saint Petersburg, 196084, Russian Federation
[b] Ulm University, Ulm, 89081, Germany
[c] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: anromanenko@corp.ifmo.ru

Article info
Received 17.01.18, accepted 20.02.18
doi: 10.17586/2226-1494-2018-18-2-350-352
Article in Russian

For citation: Romanenko A.N. Feature combination for the task of neural network acoustic model learning. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 350–352 (in Russian). doi: 10.17586/2226-1494-2018-18-2-350-352

Abstract
A method of feature combination for the problem of neural network acoustic models training is proposed aimed at the quality improvement of speech recognition. Unlike the feeding of a concatenated vector of features of a different nature to the neural network input, the proposed method uses a delayed union at the level of hidden layers. It uses individual input streams for each type of features. Such streams are able to extract specific patterns for each type of features and then combine them on the hidden layer of the neural network acoustic model. The effect of the method on the system quality was studied in the task of Russian conversational telephone speech recognition. The proposed method achieves 0.41% absolute reduction of the word error rate relative to the concatenation of features and 1.35% in comparison with the best system using one type of features. The results of the work can be used to develop automatic speech recognition systems.

Keywords
feature combination, neural network acoustic models, speech recognition

Acknowledgements
The research is supported by the Ministry of Education and Science of the Russian Federation, contract No. 8.9971.2017/DAAD

На сегодняшний день стандартом в системах распознавания речи является применение акустических моделей (АМ) на основе нейронных сетей [1, 2]. Нейросетевые АМ обеспечивают низкий уровень пословной ошибки распознавания речи (Word Error Rate, WER), который может быть в дальнейшем уменьшен за счет усреднения предсказаний ансамбля АМ [1]. Однако нейросетевые АМ обладают высокой вычислительной сложностью, что значительно снижает быстродействие систем распознавания речи [3]. Тем самым использование ансамбля моделей становится неприменимым в реальных задачах.

Альтернативным подходом для уменьшения WER может быть объединение признаков различной природы [4], которые способны учитывать разнообразные акустические особенности речевого сигнала. Так, коэффициенты перцептивного линейного предсказания (plp) характеризуют спектральную составляющую сигнала, а гамматонные фильтры (gtf) описывают фазу и частотные модуляции [5]. Комбинация частоты основного тона и вероятности вокализации (pitch) [6] также вносит значимый вклад в снижение WER. Кроме того, акустические признаки, извлеченные из так называемого узкого горла (bottleneck) [7] нейронных сетей, позволяют значительно увеличивать качество распознавания речи [8, 9].

Зачастую объединение различных признаков реализуется за счет обычной конкатенации их векторов, соответствующих одному временно́му фрагменту речевого сигнала (рисунок, а). Такой подход обладает рядом недостатков: конкатенированный вектор может обладать высокой размерностью (потребует увеличения входного слоя нейронной сети и вычислительных затрат); возможно наличие корреляции между элементами конкатенированного вектора, что негативно отражается на процессе обучения нейросетевых АМ, так как не позволяет извлечь паттерны, характерные для каждого типа признаков. Предлагается метод, лишенный данных недостатков и заключающийся в отложенном объединении признаков на уровне скрытых слоев нейронной сети. Суть метода состоит в формировании для каждого типа признаков отдельного входного потока. Каждый такой поток состоит из одного входного и одного скрытого слоя. Все входные потоки затем объединяются на скрытом слое верхнего уровня (рисунок, б). Данный подход позволяет независимо извлекать паттерны для различных признаков. Возможность варьировать размерности входных потоков независимо друг от друга позволяет подобрать оптимальную конфигурацию под каждый вид признаков.

Экспериментальная оценка применимости метода проводилась для задачи распознавания русской спонтанной речи в телефонном канале. Для обучения моделей использовался набор данных STC-train_100h (подмножество набора данных STC-train [10]) объемом 100 часов. Тестирование проводилось на наборе данных STC-test-5 [10], длительностью 3 часа 47 минут. Стоит отметить, что данный тестовый набор содержит 54 записи, которые сбалансированы по продолжительности и количеству слов. Ввиду данной особенности погрешность получаемых значений WER может быть весьма низкой.

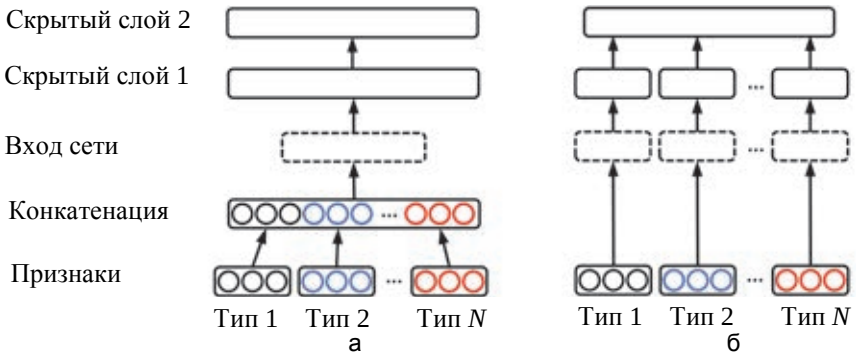

Рисунок. Нейросетевая акустическая модель с конкатенацией признаков (а) и формированием отдельных входных потоков для каждого типа признаков (б)

В ходе экспериментов для каждой конфигурации признаков было обучено по десять идентичных АМ на основе нейронной сети с запаздыванием (Time-Delayed Neural Network, TDNN) [11]. Итоговая оценка WER для каждой из конфигураций приведена в виде среднего значения и среднеквадратического отклонения десяти моделей. Для установления статистической значимости различий в WER между различными конфигурациями был применен парный t-тест Стьюдента с уровнем $p = 0{,}05$ (предварительно выборки WER были исследованы на нормальность при помощи критерия Шапиро–Уилка). Статистическая значимость была проверена как для выборок пофайлового WER, так и усредненных WER для 10 систем между различными конфигурациями. Всего было исследовано 5 конфигураций:

– TDNN1: признаки – конкатенация gtf, plp и pitch. WER=38,23±0,08%. Данная конкатенация демонстрирует высокое качество распознавания речи ввиду того, что объединяемые признаки описывают совершенно разные акустические свойства сигнала;
– TDNN2: признаки – BN2 [8]. WER=36,33±0,05%;
– TDNN3: признаки – DNN-SDBN [9]. WER=36,16±0,03%;
– TDNN4: признаки – конкатенация gtf, plp, pitch, BN2 и DNN-SDBN. WER=35,22±0,07%.
– TDNN5: признаки – BN2, DNN-SDBN и конкатенация gtf, plp, pitch. WER=34,81±0,04%.

Как видно из результатов экспериментов, конкатенация различных признаков, использованная в модели TDNN4, позволила уменьшить WER на 0,94% в сравнении с лучшей системой на одиночных признаках (TDNN3). Однако предложенный метод, использованный в TDNN5, продемонстрировал дополнительное снижение WER на 0,41% относительно TDNN4. Таким образом, предложенный метод позволил добиться 1,35% абсолютного снижения ошибки распознавания при использовании акустических признаков различной природы.

Литература

1. Siohan O., Rybach D. Multitask learning and system combination for automatic speech recognition // Proc. IEEE Workshop on Automatic Speech Recognition and Understanding. Scottsdale, USA, 2015. P. 589–595. doi: 10.1109/ASRU.2015.7404849
2. Saon G., Kurata G., Sercu T. et al. English conversational telephone speech recognition by humans and machines // Proc. INTERSPEECH. Stockholm, Sweden, 2017. P. 132–136. doi: 10.21437/Interspeech.2017-405
3. Narang S., Elsen E., Diamos G., Sengupta S. Exploring sparsity in recurrent neural networks // Proc. International Conference on Learning Representations (ICLR). Toulon, France, 2017. arXiv:1704.05119
4. Zolnay A., Schluter R., Ney H. Acoustic feature combination for robust speech recognition // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Philadelphia, USA, 2005. P. I457–I460. doi: 10.1109/ICASSP.2005.1415149
5. Pulkki V., Karjalainen M. Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics. Wiley, 2015. 454 p.
6. Ghahremani P., BabaAli B., Povey D. at al. A pitch extraction algorithm tuned for automatic speech recognition // Proc. Int. Conf. on Acoustics, Speech and Signal Processing. Florence, Italy, 2014. P. 2494–2498. doi: 10.1109/ICASSP.2014.6854049
7. Grezl F., Karafiat M., Kontar S. Probabilistic and bottle-neck features for LVCSR of meetings // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Honolulu, USA, 2007. V. 4. P. IV757–IV760. doi: 10.1109/ICASSP.2007.367023
8. Меденников И.П. Дикторо-зависимые признаки для распознавания спонтанной речи // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 1. С. 195–197. doi:10.17586/2226-1494-2016-16-1-195-197
9. Khokhlov Y., Medennikov I., Romanenko A. et al. The STC keyword search system for OpenKWS 2016 evaluation // Proc. INTERSPEECH. Stockholm, Sweden, 2017. P. 3602–3606. doi: 10.21437/Interspeech.2017-1212
10. Меденников И.П. Методы, алгоритмы и программные средства распознавания русской телефонной спонтанной речи: дис. … канд. техн. наук. СПб, 2016. 200 с.
11. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts // Proc. INTERSPEECH. Dresden, Germany, 2015. P. 3214–3218.

References

1. Siohan O., Rybach D. Multitask learning and system combination for automatic speech recognition. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. Scottsdale, USA, 2015, pp. 589–595. doi: 10.1109/ASRU.2015.7404849
2. Saon G., Kurata G., Sercu T. et al. English conversational telephone speech recognition by humans and machines. *Proc. INTERSPEECH*. Stockholm, Sweden, 2017, pp. 132–136. doi: 10.21437/Interspeech.2017-405
3. Narang S., Elsen E., Diamos G., Sengupta S. Exploring sparsity in recurrent neural networks. *Proc. International Conference on Learning Representations, ICLR*. Toulon, France, 2017. arXiv:1704.05119
4. Zolnay A., Schluter R., Ney H. Acoustic feature combination for robust speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Philadelphia, USA, 2005, pp. I457–I460. doi: 10.1109/ICASSP.2005.1415149
5. Pulkki V., Karjalainen M. *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Wiley, 2015, 454 p.
6. Ghahremani P., BabaAli B., Povey D. at al. A pitch extraction algorithm tuned for automatic speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*. Florence, Italy, 2014, pp. 2494–2498. doi: 10.1109/ICASSP.2014.6854049
7. Grezl F., Karafiat M., Kontar S. Probabilistic and bottle-neck features for LVCSR of meetings. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Honolulu, USA, 2007, vol. 4, pp. IV757–IV760. doi: 10.1109/ICASSP.2007.367023
8. Medennikov I.P. Speaker-dependent features for spontaneous speech recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 1, pp. 195–197. (In Russian) doi: 10.17586/2226-1494-2016-16-1-195-197
9. Khokhlov Y., Medennikov I., Romanenko A. et al. The STC keyword search system for OpenKWS 2016 evaluation. *Proc. INTERSPEECH*. Stockholm, Sweden, 2017, pp. 3602–3606. doi: 10.21437/Interspeech.2017-1212
10. Medennikov I.P. *Methods, Algorithms and Software for Recognition of Russian Spontaneous Phone Speech. Dis. PhD Eng. Sci*. St. Petersburg, Russia, 200 p.
11. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts. *Proc. INTERSPEECH*. Dresden, Germany, 2015, pp. 3214–3218.

Авторы

Романенко Алексей Николаевич – аспирант, научный сотрудник, ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация; аспирант, Ульмский университет, Ульм, 89081, Германия; аспирант, Университет ИТМО, 197101, Санкт-Петербург, Российская Федерация, Scopus ID: 56414341400, ORCID ID: 0000-0002-7828-968X, anromanenko@corp.ifmo.ru

Authors

Alexey N. Romanenko – postgraduate, Scientific researcher, STC Ltd., Saint Petersburg, 196084, Russian Federation; postgraduate, Ulm University, Ulm, 89081, Germany; postgraduate, ITMO University, 197101, Saint Petersburg, Russian Federation, Scopus ID: 56414341400, ORCID ID: 0000-0002-7828-968X, anromanenko@corp.ifmo.ru