НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2016 Том 16 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2016 Vol. 16 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

**УДК 004.522**

# ДИКТОРО-ЗАВИСИМЫЕ ПРИЗНАКИ ДЛЯ РАСПОЗНАВАНИЯ СПОНТАННОЙ РЕЧИ

**И.П. Меденников**[a,b]

[a] ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация
[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: ipmsbor@yandex.ru

**Информация о статье**
Поступила в редакцию 19.11.15, принята к печати 01.12.15
doi:10.17586/2226-1494-2016-16-1-195-197
Язык статьи – русский

**Ссылка для цитирования:** 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.

**Аннотация**
Приведены результаты исследования по повышению устойчивости системы распознавания спонтанной речи к акустической вариативности речевого сигнала. Предложен метод построения высокоуровневых признаков при помощи глубокой нейронной сети с узким горлом, адаптированной к диктору и акустической обстановке при помощи **i**-векторов. Предложенный метод обеспечил относительное уменьшение на 11,9% пословной ошибки в задаче распознавания русской спонтанной речи в телефонном канале.

**Ключевые слова**
автоматическое распознавание речи, адаптация к диктору, i-векторы, признаки из глубокой нейронной сети с узким горлом

# SPEAKER-DEPENDENT FEATURES FOR SPONTANEOUS SPEECH RECOGNITION

**I.P. Medennikov** [a,b]

[a] STC-Innovations Ltd., Saint Petersburg, 196084, Russian Federation
[b] ITMO University Saint Petersburg, 197101, Russian Federation
Corresponding author: ipmsbor@yandex.ru

**Article info**
Received 19.11.15, accepted 01.12.15
doi:10.17586/2226-1494-2016-16-1-195-197
Article in Russian
**For citation:** 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.

**Abstract**
This paper presents the results of the study on improving robustness to the acoustic variability of the speech signal for spontaneous speech recognition system. The method is proposed to constructing high-level bottleneck features using deep neural network adapted to a speaker and to acoustic environment with i-vectors. The proposed method provides 11,9% relative reduction of word error rate in Russian spontaneous telephone speech recognition task.

**Keywords**
automatic speech recognition, speaker adaptation, i-vectors, bottleneck features from deep neural network.

Одной из наиболее сложных задач в области автоматического распознавания речи является проблема распознавания разговорной спонтанной речи. Сложность задачи обусловлена особенностями разговорной спонтанной речи: высокие канальная и дикторская вариативность, наличие аддитивных и нелинейных искажений, наличие акцентной и эмоциональной речи, разнообразная манера произнесения, вариативность темпа речи, редукция и вялая артикуляция.

Распознаванию спонтанной речи посвящено большое количество работ, например, [1–4]. Приведенный в [4] анализ показывает, что одним из наиболее перспективных направлений повышения качества распознавания спонтанной речи является снижение чувствительности системы распознавания к акустической вариативности речевого сигнала. Эффективным способом реализации этого подхода является применение адаптации акустических моделей на основе глубоких нейронных сетей (DNN) с использованием так называемых i-векторов. Он реализует адаптацию как к диктору, так и к акустической обстановке

[5]. i-вектор представляет собой малоразмерный вектор, кодирующий отличие распределения признаков, оцененного по интересующему фрагменту речевого сигнала (как правило, это участок, соответствующий речи одного диктора), от распределения, оцененного по всей обучающей выборке. i-векторы получили широкое распространение в задаче идентификации и верификации диктора [6] и недавно были успешно применены в задачах распознавания речи [2].

Предложенный в настоящей работе метод заключается в том, чтобы извлекать высокоуровневые признаки из дикторо-зависимой глубокой нейронной сети с узким горлом, использующей i-векторы для адаптации (bottleneck-признаки), и использовать построенные признаки для обучения акустической модели. Метод состоит из следующих шагов.

Шаг 1. Обучение глубокой нейронной сети, адаптированной при помощи i-векторов.

Шаг 2. Разбиение последнего скрытого слоя на два слоя следующим образом:

$$\mathbf{y} = f(\mathbf{W}\mathbf{x} + \mathbf{b}) \approx f(\mathbf{W}_{out}(\mathbf{W}_{bn}\mathbf{x}) + \mathbf{b}),$$

где $f$, $\mathbf{y}$, $\mathbf{W}$ и $\mathbf{b}$ – соответственно функция активации, вектор активации**,** матрица весов и вектор смещений разбиваемого слоя глубокой нейронной сети; $\mathbf{x}$ – вектор активации слоя, предшествующего разбиваемому; $\mathbf{W}_{bn}$ – матрица весов малоразмерного слоя (называемого также узким горлом или bottleneck-слоем) с линейной функцией активации; $\mathbf{W}_{out}$ – матрица весов нелинейного слоя, имеющего размерность исходного разбиваемого слоя. Разбиение $\mathbf{W} \approx \mathbf{W}_{out}\mathbf{W}_{bn}$ осуществляется при помощи сингулярного разложения матрицы $\mathbf{W}$.

Шаг 3. Дообучение полученной глубокой нейронной сети с узким горлом по критерию минимизации взаимной энтропии с использованием $L_2$-штрафа на отклонение параметров от исходных значений.

Шаг 4. Отбрасывание слоев глубокой нейронной сети, следующих за узким горлом, и использование полученной нейронной сети для построения признаков.

Предполагается, что чем лучше базовая нейронная сеть, тем лучше будут получаемые из нее bottleneck-признаки. Исходя из этого, схема была адаптирована для применения с глубокими нейронными сетями, обученными по критерию разделения последовательностей (sequence-discriminative training, [1]). Прямолинейное решение, заключающееся в использовании критерия разделения последовательностей на шаге 3, не дало улучшения. В связи с этим был предложен альтернативный способ – обучение по критерию минимизации взаимной энтропии с использованием в качестве эталонных вероятностей, генерируемых исходной моделью, обученной по критерию разделения последовательностей.

Обучение акустических моделей осуществлялось при помощи набора инструментов Kaldi ASR[1] [1, 7] на базе русской телефонной спонтанной речи объемом 390 ч, содержащей записи нескольких сотен дикторов.

В качестве входных признаков для базовой конфигурации DNN использовались логарифмы мощности выходов 20-ти треугольных мел-частотных фильтров (FBANK) с нормализацией среднего спектра, взятые с временным контекстом в 31 кадр. Выходной софтмакс-слой содержал 13000 нейронов, соответствующих состояниям трифонной акустической модели на основе гауссовых смесей. Наилучшее качество было достигнуто при использовании 6 скрытых слоев по 1024 нейрона в каждом, с сигмоидами в качестве функций активации. Обучение адаптированной DNN проводилось с i-векторами размерности 50 по схеме, предложенной в работе [8], состоящей из следующих шагов.

Шаг 1. Обучение дикторо-независимой DNN на исходных признаках.

Шаг 2. Расширение входного слоя обученной на шаге 1 дикторо-независимой DNN.

Шаг 3. Дообучение DNN с расширенным входным слоем с использованием $L_2$-штрафа на отклонение параметров сети от исходной модели.

Векторы дикторо-зависимых bottleneck-признаков размерности 80 были построены при помощи DNN, обученных на FBANK+i-вектор признаках по критерию минимизации взаимной энтропии и по критерию разделения последовательностей sMBR [1] (BN1 и BN2 соответственно). Обучение DNN с использованием построенных дикторо-зависимых bottleneck-признаков осуществлялось с временным контекстом в 31 кадр с шагом в 5 кадров, аналогично работе [9]. Лучшие результаты продемонстрировала конфигурация DNN с 4 скрытыми слоями по 2048 нейронов с сигмоидами в качестве функций активации. Также было выполнено несколько итераций дообучения DNN по sMBR-критерию.

Тестирование проводилось на записях телефонных переговоров на русском языке продолжительностью около 1 ч, принадлежащих 9 различным дикторам. Использовалась триграммная языковая модель со словарем в 200 тысяч слов. Коэффициент неопределенности на тестовых данных составил 360, частотность внесловарных слов составила 1,8%. Для оценки эффективности предложенного метода было проведено распознавание тестовой базы с использованием набора инструментов Kaldi ASR и четырех акустических моделей на основе DNN, обученных на различных признаках по sMBR-критерию. Неадаптированная акустическая модель на FBANK-признаках принята за базовую для сравнения. Значения пословной ошибки распознавания (WER), а также абсолютного ($\Delta WER = WER_1 - WER_2$) и относительного

[1] http://www.kaldi-asr.org

($WERR = (WER_1 - WER_2)/WER_1$) уменьшения пословной ошибки распознавания представлены в таблице. Они свидетельствуют о том, что адаптация DNN с использованием i-векторов позволяет значительно улучшить точность распознавания речи. Предложенный в работе метод построения высокоуровневых дикторо-зависимых признаков позволяет получить дополнительное улучшение точности распознавания речи.

| Акустическая модель (признаки) | WER, % | ΔWER, % | WERR, % |
|---|---|---|---|
| Неадаптированная (FBANK) | 28,5 | – | – |
| Адаптированная (FBANK+**i**-вектор) | 26,0 | 2,5 | 8,6 |
| Адаптированная (BN1) | 25,3 | 3,2 | 11,2 |
| Адаптированная (BN2) | 25,1 | 3,4 | 11,9 |

Таблица. Результаты экспериментов

**References**


1. Vesely K., Ghoshal A., Burget L., Povey D. Sequence-discriminative training of deep neural networks. *Proc. of the Annual Conference of International Speech Communication Association, INTERSPEECH*. Lyon, France, 2013, pp. 2345–2349.
2. Saon G., Soltau H., Nahamoo D., Picheny M. Speaker adaptation of neural network acoustic models using i-vectors. *Proc. IEEE workshop on Automatic Speech Recognition and Understanding, ASRU*. Olomouc, Czech Republic, 2013, pp. 55–59. doi: 10.1109/ASRU.2013.6707705
3. Soltau H., Saon G., Sainath T.N. Joint training of convolutional and non-convolutional neural networks. *Proc. International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Florence, Italy, 2014, pp. 5572–5576. doi: 10.1109/ICASSP.2014.6854669
4. Prudnikov A., Medennikov I., Mendelev V., Korenevsky M., Khokhlov Y. Improving acoustic models for Russian spontaneous speech recognition. *Lecture Notes in Computer Science*, 2015, vol. 9319, pp. 234–242. doi: 10.1007/978-3-319-23132-7_29
5. Rouvier M., Favre B. Speaker adaptation of DNN-based ASR with i-vectors: does it actually adapt models to speakers? *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH*. Singapore, 2014, pp. 3007–3011.
6. Kozlov A., Kudashev O., Matveev Y., Pekhovsky T., Simonchik K., Shulipa A. SVID speaker recognition system for NIST SRE 2012. *Lecture Notes in Computer Science*. Pilsen, Czech Republic, 2013, vol. 8113, pp. 278–285. doi: 10.1007/978-3-319-01931-4_37
7. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. The Kaldi speech recognition toolkit. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*. Waikoloa, USA, 2011, pp. 1–4.
8. Senior A., Lopez-Moreno I. Improving DNN speaker independence with I-vector inputs. *Proc. International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Florence, Italy, 2014, pp. 225–229. doi: 10.1109/ICASSP.2014.6853591
9. Karafiat M., Grezl F., Hannemann M., Cernocky J. But neural network features for spontaneous Vietnamese in BABEL. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Florence, Italy, 2014, pp. 5622–5626. doi: 10.1109/ICASSP.2014.6854679



***Меденников Иван Павлович*** – научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ipmsbor@yandex.ru

***Ivan P. Medennikov*** – scientific researcher, STC-Innovations Ltd., Saint Petersburg, 196084, Russian Federation; engineer, ITMO University Saint Petersburg, 197101, Russian Federation, ipmsbor@yandex.ru