НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
нотябрь–декабрь 2016 Том 16 № 6 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2016 Vol. 16 No 6 ISSN 2226-1494 http://ntv.ifmo.ru/en
УНИВЕРСИТЕТ ИТМО

УДК 004.934

ИСПОЛЬЗОВАНИЕ В СИСТЕМАХ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ GMM-МОДЕЛЕЙ ДЛЯ АДАПТАЦИИ АКУСТИЧЕСКИХ МОДЕЛЕЙ, ПОСТРОЕННЫХ НА ОСНОВЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Н.А. Томашенко[a,b,c], Ю.Ю. Хохлов[b], Э. Ларшер[a], Я. Эстев[a], Ю.Н. Матвеев[b,c]

[a] Лаборатория Информатики Университета Ле Мана (LIUM), Ле Ман, 72085, Франция
[b] ООО «ЦРТ-Инновации», Санкт-Петербург, 196084, Российская Федерация
[c] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: tomashenko-n@speechpro.com

Информация о статье
Поступила в редакцию 04.10.16, принята к печати 30.10.16
doi: 10.17586/2226-1494-2016-16-6-1063-1072
Язык статьи – русский

Ссылка для цитирования: Томашенко Н.А., Хохлов Ю.Ю., Ларшер Э., Эстев Я., Матвеев Ю.Н. Использование в системах автоматического распознавания речи GMM-моделей для адаптации акустических моделей, построенных на основе искусственных нейронных сетей // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 6. С. 1063–1072. doi: 10.17586/2226-1494-2016-16-6-1063-1072

Аннотация

Предмет исследования. Исследованы вопросы адаптации к диктору акустических моделей, построенных на основе искусственных нейронных сетей, для задачи автоматического распознавания речи. Цель адаптации к диктору заключается в улучшении точности работы системы автоматического распознавания речи при работе с конкретным диктором. **Метод.** Метод обучения и адаптации акустических моделей на основе глубоких нейронных сетей использует вспомогательную GMM (Gaussian Mixture Models, модель смеси гауссовских распределений) и GMMD (GMM-derived, полученные с использованием GMM) признаки. Главное достоинство предложенных GMMD-признаков состоит в возможности адаптации DNN (Deep Neural Network, глубокая нейронная сеть) модели посредством адаптации вспомогательной GMM-модели. Предложенный подход позволяет применять любые алгоритмы адаптации GMM для адаптации DNN-моделей и является универсальным способом переноса адаптационных техник из фреймворка GMM во фреймворк DNN-моделей. **Основные результаты.** Эффективность работы предлагаемого подхода проверена с использованием одного из наиболее распространенных алгоритмов адаптации GMM-моделей – MAP (Maximum A Posteriori) адаптации. Предложены и изучены разные способы интеграции предлагаемого подхода в современную архитектуру нейросетевых акустических моделей. Проведен анализ выбора типа GMM. Результаты экспериментов на корпусе TED-LIUM показали эффективность предложенного подхода: в режиме адаптации без учителя предложенный алгоритм адаптации и рассмотренные методы фьюжена позволяют достичь 11–18% относительного уменьшения пословной ошибки распознавания по сравнению с дикторо-независимой акустической моделью, построенной по традиционному рецепту на стандартных признаках, и на 3–6% – по сравнению с дикторо-адаптированной базовой моделью.

Ключевые слова
автоматическое распознавание речи, акустические модели, адаптация к диктору, глубокие нейронные сети, GMMD-признаки, MAP, fMLLR, GMM, адаптация акустических моделей, фьюжен

Благодарности
Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

GAUSSIAN MIXTURE MODELS FOR ADAPTATION OF DEEP NEURAL NETWORK ACOUSTIC MODELS IN AUTOMATIC SPEECH RECOGNITION SYSTEMS

N.A. Tomashenko[a,b,c], Yu.Yu. Khokhlov[b], A. Larcher[a], Ya. Estève[a], Yu.N. Matveev[b,c]

[a] Laboratory of Computer Science of the University of Le Mans (LIUM), Le Mans, 72085, France
[b] "STC -Innovation" Ltd., Saint Petersburg, 196084, Russian Federation
[c] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: tomashenko-n@speechpro.com

Article info
Received 04.10.16, accepted 30.10.16
doi: 10.17586/2226-1494-2016-16-6-1063-1072
Article in Russian
For citation: Tomashenko N.A., Khokhlov Yu.Yu., Larcher A., Estève Ya., Matveev Yu. N. Gaussian mixture models for adaptation of deep neural network acoustic models in automatic speech recognition systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 6, pp. 1063–1072. doi: 10.17586/2226-1494-2016-16-6-1063-1072

Abstract
Subject of Research. We study speaker adaptation of deep neural network (DNN) acoustic models in automatic speech recognition systems. The aim of speaker adaptation techniques is to improve the accuracy of the speech recognition system for a particular speaker. **Method**. A novel method for training and adaptation of deep neural network acoustic models has been developed. It is based on using an auxiliary GMM (Gaussian Mixture Models) model and GMMD (GMM-derived) features. The principle advantage of the proposed GMMD features is the possibility of performing the adaptation of a DNN through the adaptation of the auxiliary GMM. In the proposed approach any methods for the adaptation of the auxiliary GMM can be used, hence, it provides a universal method for transferring adaptation algorithms developed for GMMs to DNN adaptation. **Main Results.** The effectiveness of the proposed approach was shown by means of one of the most common adaptation algorithms for GMM models – MAP (Maximum A Posteriori) adaptation. Different ways of integration of the proposed approach into state-of-the-art DNN architecture have been proposed and explored. Analysis of choosing the type of the auxiliary GMM model is given. Experimental results on the TED-LIUM corpus demonstrate that, in an unsupervised adaptation mode, the proposed adaptation technique can provide, approximately, a 11–18% relative word error reduction (WER) on different adaptation sets, compared to the speaker-independent DNN system built on conventional features, and a 3–6% relative WER reduction compared to the SAT-DNN trained on fMLLR adapted features.

Keywords
automatic speech recognition (ASR), acoustic models, speaker adaptation, deep neural networks (DNN), GMM-derived features, GMMD, maximum a posteriori (MAP), fMLLR, GMM, acoustic model adaptation, fusion
Acknowledgements
The work is partially financially supported by the Government of the Russian Federation (grant 074-U01).

Введение

Успех в исследованиях и разработке систем автоматического распознавания речи привел к появлению большого разнообразия приложений, таких как интерактивные диалоговые системы и системы информационного поиска, автоматическое субтитрование телевизионных передач, системы диктовки, распознавания записанной речи, системы обучения языкам и многие другие. Однако до сих пор возникают некоторые сложности, когда приложение системы автоматического распознавания речи (APP) используется в условиях, отличных от тех, в которых эта система была обучена. Различия в условиях обучения системы APP и условиях ее использования в реальности могут привести к ухудшению качества работы этой системы и снижению точности распознавания речи. Причины различий могут быть обусловлены разными факторами, такими как тип канала связи, наличие шумов в окружении говорящего, особенностями речи диктора. Адаптация является эффективным способом уменьшения несоответствий между условиями обучения и условиями использования системы APP. Целью адаптации к диктору является улучшение качества распознавания целевого диктора, речь которого требуется распознать в реальных (тестовых) условиях, при использовании небольшого количества звуковых данных этого диктора.

За последние несколько лет в современных системах APP широкое распространение получили гибридные акустические модели, основанные на использовании глубоких нейронных сетей (Deep Neural Network, DNN) и скрытых марковских моделей (Hidden Markov Models, HMM), поскольку было показано, что во многих задачах они значительно превосходят акустические модели, основанные на смеси гауссовых распределений (GMM) [1]. В связи с этим фактом задача адаптации акустических моделей снова приобрела чрезвычайную актуальность, так как хорошо изученные методы адаптации, разработанные в прошлом для GMM [2, 3], не могут быть напрямую использованы для адаптации DNN-моделей из-за разной природы этих типов моделей.

Большое количество параметров в гибридных акустических моделях делает очень сложной задачу адаптации, особенно в условиях ограниченного количества адаптационных данных, и требует разработки новых подходов. При адаптации моделей с таким большим количеством параметров, как в современных DNN-моделях, необходимо особенно внимательно подходить к проблеме переобучения моделей.

Среди современных методов адаптации акустических моделей, построенных на основе нейронных сетей, можно выделить следующие классы:

– *методы, использующие линейные преобразования*, которые могут быть выполнены на разных уровнях нейронной сети, такие как LIN [4] (Linear Input Network transformation – линейное преобразование входного слова) и fDLR [5] (feature space Discriminative Linear Regression – дискриминативная линейная регрессия в пространстве признаков), где линейное преобразование применяется к входным признакам, на которых обучается нейронная сеть. В LHN [4] (Linear Hidden Network transformation – линейное преобразование скрытого слова) линейное преобразование применяется к функциям активации скрытых слоев, а в LON (Linear Output Network transformation – линейное преобразование выход-

ного слоя сети) и oDLR [6] (output-feature Discriminative Linear Regression – дискриминативная линейная регрессия выходных признаков) применяется к выходному (softmax) слою;
- дообучение нейронной сети или только ее части с использованием *регуляризации*, например, L2-prior регуляризации [7] или регуляризации на основе расстояния Кульбака–Лейблера [8];
- применение концепции *многозадачного* (*multi-task*) *обучения* [9–12];
- *использование вспомогательных признаков*, примерами которых могут служить i-вектора (i-vectors или identity vectors – вектора идентичности дикторов) [13–16] и коды дикторов [17];
- *совместное использование GMM- и DNN-моделей* [18–29]. Наиболее распространенный способ комбинации GMM и DNN для адаптации заключается в использовании признаков, адаптированных с помощью GMM, например, fMLLR, в качестве входа для обучения DNN-модели [18–20]. К другим методам этого типа относятся temporally varying weight regression [21] и GMMD-признаки [23–29].

Методы адаптации GMM имеют гораздо более долгую историю, чем методы адаптации DNN, хорошо изучены и доказали свою эффективность в разных условиях применения. Среди вышеперечисленных методов для DNN только последняя группа использует адаптивные свойства GMM-моделей в решении задачи адаптации DNN-моделей. Однако среди алгоритмов [18–21] нет универсального метода, позволяющего перенести все алгоритмы адаптации GMM-моделей на работу с DNN-моделями. Цель данной работы – сделать шаг в этом направлении, разработать такой подход, который бы позволил применить для адаптации DNN-моделей широкий класс алгоритмов, разработанный в прошлом для GMM-моделей.

Адаптация к диктору с использованием GMMD-признаков

Использование GMMD-признаков для адаптации нейронных сетей впервые было предложено в [23], где на примере MAP (Maximum a Posteriori – максимум апостериорной вероятности) адаптации была показана эффективность данного подхода. В дальнейшем [24–28] данный подход был исследован для других тестовых корпусов, других топологий акустических моделей и других алгоритмов адаптации, например, для fMLLR (feature space Maximum Likelihood Linear Regression – максимум правдоподобия линейной регрессии в пространстве признаков).

Основная идея предложенного в настоящей работе подхода заключается в специальном способе вычисления признаков, на которых строятся гибридные акустические модели. Так называемые GMMD-признаки вычисляются с использованием вспомогательной GMM-модели (из оценок правдоподобий состояний монофонов или трифонов). Предлагаемый подход построения входных признаков для обучения DNN позволяет выполнять адаптацию к диктору посредством адаптации вспомогательной GMM, использованной для построения признаков. Для адаптации вспомогательной GMM в этом случае может быть применен любой существующий алгоритм адаптации GMM, например, MLLR [2] или MAP-адаптация [3], поэтому можно сказать, что такой подход представляет GMM-фреймворк для адаптации DNN-моделей.

Предлагаемая схема обучения акустических моделей с использованием GMMD-признаков показана на рис. 1 и состоит из следующих шагов.

1. Извлечение из речевого сигнала акустических векторов признаков. Это могут быть, например, 39-размерные мел-частотные кепстральные коэффициенты (Mel Frequency Cepstral Coefficients, MFCC) с первыми и вторыми производными, как в работах [24, 25], либо bottleneck (BN) признаки, как в работах [26, 28]. Признаки извлекаются с шагом 10 мс и размером скользящего окна 25 мс.
2. Нормализация полученных признаков (Cepstral Mean Normalization, CMN).
3. Полученные в п. 2 признаки вместе с подаваемыми на вход транскрибированными текстовыми файлами используются для адаптации вспомогательной GMM-модели. Важно отметить, что на данном этапе можно использовать любой алгоритм адаптации GMM-HMM. В данном исследовании были проведены эксперименты для одного из наиболее распространенных алгоритмов адаптации GMM-HMM-моделей – MAP-адаптация. Вспомогательная GMM-HMM состоит из моделей монофонов или трифонов, каждый из которых содержит несколько состояний.
4. Адаптированная к диктору в п. 3 GMM-модель используется для преобразования признаков, полученных в п. 2, в вектора значений логарифмов вероятностей. Обозначим $\mathbf{o}_t$ – акустический вектор признаков в момент времени t, тогда новый GMMD-вектор признаков $\mathbf{f}_t$ вычисляется следующим образом:
$$\mathbf{f}_t = [p_t^1, \dots, p_t^n], \quad (1)$$
где n – это количество состояний во вспомогательной монофонной GMM-модели, а
$$p_t^i = \log(P(\mathbf{o}_t | s_t = i)) \quad (2)$$
есть логарифм вероятности, оцениваемый с помощью GMM-модели; s_t – индекс состояния в момент времени t. Данная процедура приводит к формированию вектора признаков размерности n для каждого акустического вектора признаков $\mathbf{o}_t$.
5. Увеличение контекста признаков, полученных в п. 4, за счет объединения нескольких соседних векторов в супервектор большей размерности путем их конкатенации.

6. Обучение DNN-HMM-модели на полученных в п. 5 признаках.
   В п. 4 были получены дикторо-зависимые GMMD-признаки, т.е. DNN-модель строится на дикторо-нормализованном пространстве признаков.

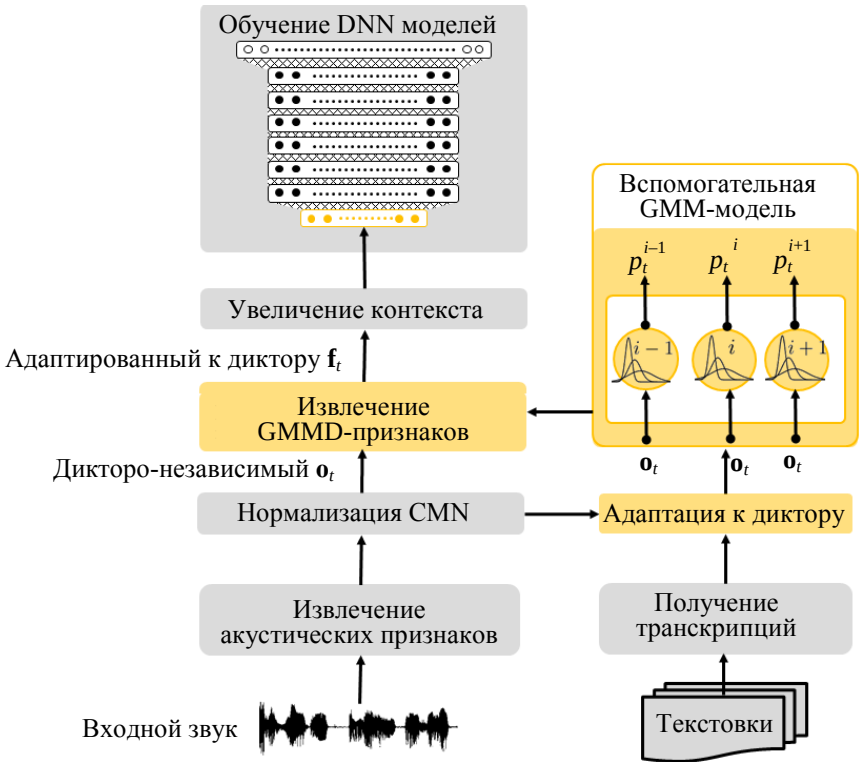

Рис. 1. Схема обучения акустических моделей на основе DNN с использованием GMMD-признаков

Выбор типа вспомогательной GMM

Вспомогательная GMM, как описано выше, используется для извлечения GMMD-признаков. От этой модели зависит как качество самих признаков, так и качество адаптации. На начальном этапе исследования [23] мы использовали монофонную GMM, в которой смеси гауссовых распределений соответствуют состояниям монофов (фонем, без учета контекста). Если каждый монофон моделируется тремя состояниями, то размерность GMMD-вектора признаков n будет равна $3 \times m$, где m – это количество монофонов в вспомогательной GMM. Известно, что трифонные GMM, состоящие из состояний трифонов (фонем, с учетом правого и левого контекста) при достаточном количестве данных для обучения превосходят монофонные GMM в качестве распознавания, поскольку позволяют более точно моделировать акустические данные с учетом контекста. Это служит мотивацией к тому, чтобы перейти в рассматриваемом методе от монофонной GMM к трифонной.

Однако переход к трифонной GMM имеет ряд сложностей и ограничений, которые мы постарались здесь проанализировать и учесть. Во-первых, при переходе к трифонной акустической модели значительно (обычно в сотни или тысячи раз) возрастает количество классов (состояний трифонов), а следовательно, увеличивается и размерность GMMD-вектора признаков, что особенно критично из-за того, что для обучения DNN используются векторы с увеличенным контекстом. Во-вторых, при увеличении количества классов может возникнуть проблема на этапе адаптации – есть вероятность, что для каких-то состояний трифонов не будет данных в адаптационной выборке, и они могут оказаться неадаптированными. Особенно это скажется для небольших адаптационных выборок.

Мы рассмотрели следующие возможные пути использования трифонных GMM в данной задаче и сравнили их эффективность.

1. Применение традиционных методов понижения размерности признаков, таких как анализ главных компонент (Principal Component Analysis, PCA).
2. Использование вместо трех значений от каждого трифона, соответствующих трем состояниям этого трифона, только максимального из этих значений. Такой подход позволяет более точно моделировать классы, а потом сократить размерность вектора в три раза.
3. Использование только наиболее частотных состояний трифонов. Иначе говоря, сначала строится обычная трифонная GMM. Потом по обучающей выборке оценивается частотность каждого трифона и для вычисления GMMD-признаков используется только заданное количество наиболее частотных трифонов.

4. Построение bottleneck-признаков небольшой размерности на основе GMMD как промежуточный этап обучения системы.

Предварительные эксперименты показали, что два последних подхода оказались наиболее эффективными для решения поставленной задачи. Дальнейшие эксперименты задействуют последнюю из этих идей.

Результаты экспериментов

Эксперименты проводились на общедоступной базе фонограмм TED-LIUM [30]. Корпус содержит 1495 лекций на английском языке с сайта TED (Technology, Entertainment, Design)[1], соответствующих 207 часам аудиоданных (141 ч – мужских голосов и 66 ч – женских) от 1242 дикторов, записанных в микрофонном канале, частота дискретизации – 16 кГц. Для экспериментов с адаптацией мы сформировали из исходного корпуса 4 подмножества: для обучения (171,7 ч, 1029 дикторов), для настройки параметров обучения (Dev: 3,5 ч, 14 дикторов) и два для тестирования качества алгоритма адаптации ($Test_1$: 3,5 ч, 14 дикторов; $Test_2$: 4,9 ч, 14 дикторов). Более подробно описание характеристик и состава этих корпусов приведено в работе [26].

Во всех экспериментах по распознаванию речи был использован словарь из 150000 слов и общедоступная тригамная языковая модель *cantab-TEDLIUMpruned.lm3* (http://cantabresearch.com/cantab-TEDLIUM.tar.bz2).

Базовая система. Для акустических моделей был использован инструментарий Kaldi toolkit [31]. На описанных выше данных мы сначала обучили базовую систему, следуя современному рецепту из Kaldi для корпуса TED-LIUM. Базовая система используется для сравнительного анализа качества работы предложенного подхода. Для обучения базовой DNN-модели сначала на 39-размерных акустических признаках (13 MFCC, плюс их первые и вторые производные) [32] была построена исходная GMM. Далее новая GMM на адаптированных с помощью fMLLR-акустических признаков была построена после применения линейного дискриминативного анализа (Linear Discriminant Analysis, LDA) и MLLT (Maximum Likelihood Linear Transform). На основе этой модели было проведено дискриминативное обучение с критерием BMMI (Boosted Maximum Mutual Information) [33]. Далее полученная GMM используется для построения DNN-модели: с помощью этой GMM-модели проводится автоматическая сегментация звуковых фонограмм по имеющимся транскрипциям [34] и трифонное связывание фонем этой модели для последующего обучения DNN. При этом сначала строится первая DNN для извлечения 40-размерных BN-признаков [28]. Входными признаками для обучения этой DNN являются 40-размерные log-scale filterbank признаки (логарифмы мощностей выходов треугольных выходов, вычисленные по шкале Мэла), конкатенированные с 3-размерными признаками на основе основного тона. Полученные 43-размерные векторы признаков перед обучением конкатенируются по 11 соседних векторов (5 справа и 5 слева от текущего вектора) в супервекторы размерности $43 \times 11 = 473$, размерность которых затем уменьшается до 258 с помощью дискретного косинусного преобразования (Discrete Cosine Transform, DCT). Таким образом, при обучении DNN для извлечения BN-признаков используется следующая топология нейронной сети: 258-размерный входной слой; 4 скрытых слоя, в которых третий слой является BN-слоем размерности 40, а остальные 3 слоя имеют размерность 1500; 2390-размерный выходной слой. На полученных BN-признаках, к которым далее применяется fMLLR-адаптация и увеличение размерности и контекста путем конкатенации, строится новая DNN-модель (базовая) со следующей топологией нейронной сети: 520-размерный входной слой; 6 скрытых слоя размерности 2048; 4184-размерный выходной слой (softmax), соответствующий состояниям трифонов. Параметры DNN инициализировались с помощью RBM (restricted Boltzmann machines – ограниченные машины Больцмана) путем послойного предобучения и обучались по кросс-энтропийному критерию, после чего применялось несколько эпох дискриминативного дообучения с sMBR (state Minimum Bayes Risk – минимум байесовского риска, вычисляемый по состояниям) критерием. Таким образом, мы обучили 2 акустические DNN-HMM-модели: одну – как описано выше (это базовая модель с адаптацией fMLLR к диктору), и вторую – точно такую же модель, но без адаптации. Вторая модель нужна только для сравнительного анализа влияния адаптации на качество распознавания речи. Более подробное описание построения акустических моделей можно найти в работе [28].

Предложенный подход с GMMD-признаками и MAP-адаптацией. Далее аналогичным образом были построены акустические DNN-модели, но с использованием предложенных GMMD-признаков и предлагаемого подхода к адаптации. Целью данных экспериментов было предложить, изучить и выбрать оптимальный способ использования GMMD-признаков и алгоритма адаптации, а также возможность его эффективной интеграции в современный рецепт обучения акустических моделей. Для этой цели мы обучили две акустические DNN-модели с разными входными признаками, как показано на рис. 2. Номера DNN-моделей на рисунке соответствуют номерам моделей в табл. 1 и 2. При построении признаков типа №1 использовалась монофонная вспомогательная GMM, а при построении признаков типа №2 – трифонная.

[1] https://www.ted.com

Результаты адаптации приведены в табл. 1 в терминах пословной ошибки распознавания (word error rate, WER), вычисляемой как умноженное на 100 отношение общего количества ошибок распознавания (замен, вставок и пропусков) к общему количеству слов в исходном тексте, который был произнесен [35].

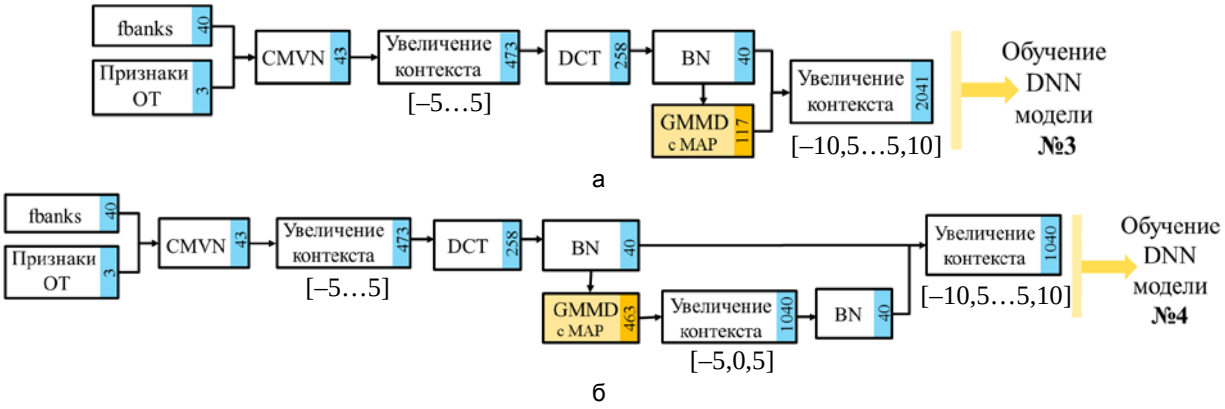

Рис. 2. Схема вычисления признаков с использованием GMMD-признаков и MAP-адаптации для обучения DNN-моделей: тип №1 (а); тип №2 (б). (ОТ – основной тон, fbanks – filterbank признаки)

Номер модели	Акустическая модель	WER, %		
		Dev	$Test_1$	$Test_2$
1	Базовая, без адаптации	12,14	10,77	13,75
2	Базовая, с fMLLR-адаптацией	10,64 (10,57)	9,52 (9,46)	12,78 (12,67)
3	На GMMD-признаках (тип №1)	10,26 (10,23)	9,40 (9,31)	12,52 (12,46)
4	На GMMD-признаках (тип №2)	10,42 (10,37)	9,74 (9,69)	13,29 (13,23)

Таблица 1. Результаты распознавания для базовых акустических моделей (номера 1 и 2) и для акустических моделей, построенных с использованием предложенного подхода (номера 3 и 4). В скобках указаны результаты распознавания из consensus-гипотез

Результаты объединения (фьюжена) базовой адаптированной модели и моделей, построенных с использованием адаптированных GMMD-признаков, даны в табл. 2. Признаки GMMD и традиционные акустические признаки имеют разную природу, вследствие чего несут взаимодополняющую информацию о речевом сигнале. В этой связи фьюжен таких моделей может улучшить качество распознавания речи – это послужило мотивацией к данным экспериментам. Мы исследовали два типа фьюжена. Первый тип – фьюжен на уровне выходов нейронных сетей (апостериорных вероятностей) акустических моделей. Второй тип фьюжена осуществлялся на более высоком уровне систем распознавания и состоит в объединении результатов распознавания в виде сеток декодирования от разных акустических моделей в единую сеть по методике CNC (confusion network combination – объединение сетей спутывания) [36]. Коэффициент **α** в таблице – это вес базовой дикторо-адаптированной модели (№2) в комбинации, оптимальное значение **α** было найдено на множестве Dev.

Номер модели	Акустические модели	**α**	WER, %, ↓ΔWER, %		
			Dev	$Test_1$	$Test_2$
5	Фьюжен апостериоров от моделей 2 и 3	0,45	9,91 ↓**6,2**	9,06 ↓**4,3**	12,04 ↓**5,0**
6	Фьюжен апостериоров от моделей 2 и 4	0,55	9,91 ↓**6,2**	9,10 ↓3,8	12,23 ↓3,5
7	Фьюжен сеток от моделей 2 и 3	0,44	10,06 ↓4,8	9,09 ↓4,0	12,12 ↓4,4
8	Фьюжен сеток от моделей 2 и 4	0,50	10,01 ↓5,3	9,17 ↓3,1	12,25 ↓3,3

Таблица 2. Результаты фьюжена базовой акустической модели (номер 2) и акустических моделей, построенных с использованием предложенного подхода (номера 3 и 4) для двух типов фьюжена – на уровне апостериорных вероятностей (номера 5 и 6) и на уровне сеток распознавания (номера 7 и 8). Здесь ΔWER, % – значения, отмеченные синим цветом после знака «↓» – относительное уменьшение пословной ошибки (если сравнивать с базовой адаптированной моделью). Жирным шрифтом выделены лучшие результаты

Визуализация признаков. Для лучшего понимания того, как ведут себя предложенные GMMD-признаки при MAP-адаптации и при обучении нейронной сети по сравнению с базовыми BN-признаками, мы использовали визуализацию входных векторов признаков и векторов выходов нейронных сетей (softmax слоя) с помощью алгоритма t-SNE (t-Distributed Stochastic Neighbor Embedding, метода стохастического вложения соседей с распределением Стьюдента) [37]. Этот алгоритм позволяет визуализировать данные большой размерности путем отображения их в двумерное или трехмерное про-

странство таким образом, что векторы, которые были близки в исходном многомерном пространстве признаков, будут близкими в новом пространстве низкой размерности. Для этого эксперимента были взяты данные пяти дикторов и акустические модели, построенные на трех типах признаков: (1) дикторо-независимые векторы BN-признаков; (2) векторы BN-признаков, адаптированные с помощью fMLLR; (3) векторы признаков GMMD, адаптированные с помощью MAP. Каждому из перечисленных типов признаков соответствует акустическая модель (нейронная сеть), обученная на этом типе признаков, как описано выше. Мы выбрали 7 разных фонем английского языка для наглядности: /ɑ/, /ɛ/, /n/, /r/, /t/, /p/, /ʃ/. Для визуального анализа по заданной сегментации мы брали только те векторы, которые относятся к центральным состояниям моделей (трифонов). Результаты для данных, посчитанных для трех моделей, показаны на рис. 3. Разным фонемам соответствуют разные цвета на рисунке. Видно, что оба вида адаптации – fMLLR на BN-признаках (рис. 3, б) и MAP на GMMD-признаках (рис. 3, в) – позволяют получить более компактные и разделимые кластеры по сравнению с дикторо-независимой моделью (рис. 3, а). При этом классы для адаптированных GMMD-признаков выглядят более компактными и разделимыми по сравнению с адаптированными BN-признаками.

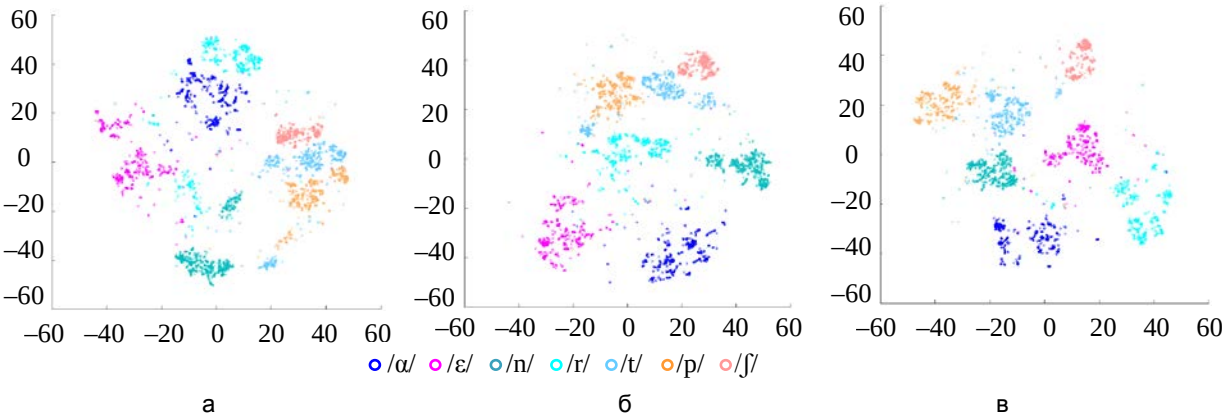

Рис. 3. Визуализация выходов softmax слоя трех нейронных сетей для 7 фонем: для дикторо-независимой модели, построенной на BN-признаках (а); для модели, построенной на BN-признаках с fMLLR-адаптацией (б); для модели, построенной на GMMD-признаках с MAP-адаптацией (в)

Заключение

В работе был исследован предложенный GMM-фреймворк для адаптации DNN-акустических моделей, а также комбинация традиционных и предложенных GMM-derived признаков для задачи адаптации к диктору DNN-HMM-акустических моделей на разных уровнях архитектуры обучения глубоких нейронных сетей. Результаты экспериментов на TED-LIUM-корпусе показали, что в режиме адаптации без учителя предложенный алгоритм адаптации и техники фьюжена позволяют достичь 11–18% относительного уменьшения пословной ошибки распознавания по сравнению с дикторо-независимой акустической моделью, построенной по традиционному рецепту на стандартных признаках, и 3–6% – по сравнению с дикторо-адаптированной базовой моделью, использующей fMLLR-адаптацию. Эксперименты с разными видами фьюжена показали, что MAP-адаптация на GMMD-признаках может дополнять fMLLR-адаптацию на традиционных BN-признаках. Оба вида фьюжена – на уровне апостериорных вероятности и на уровне сеток декодирования – дают похожий дополнительный прирост в качестве распознавания, при этом фьюжен на уровне апостериорных вероятностей оказался немного эффективнее, чем фьюжен на уровне сеток.

Дальнейшее направление исследования проблемы состоит в усовершенствовании процедуры обучения акустических моделей на GMMD-признаках, в исследовании возможности совмещения других адаптационных техник с предложенным подходом, а также в применении предложенной идеи для нейросетевых акустических моделей новых типов, таких как рекуррентные нейронные сети (Recurrent Neural Networks, RNN), нейронные сети долгой кратковременной памяти (Long Short-Term Memory, LSTM) и другие.

Литература

1. Hinton G., Deng L., Yu D., Dahl G., Mohamed A.-R., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups // IEEE Signal Processing Magazine. 2012. V. 29. N 6. P. 82–97. doi: 10.1109/MSP.2012.2205597
2. Gales M.J. Maximum likelihood linear transformations for

References

1. Hinton G., Deng L., Yu D., Dahl G., Mohamed A.-R., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, vol. 29, no. 6, pp. 82–97. doi: 10.1109/MSP.2012.2205597
2. Gales M.J. Maximum likelihood linear transformations for

HMM-based speech recognition // Computer Speech and Language. 1998. V. 12. N 2. P. 75–98.
3. Gauvain J.-L., Lee C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains // IEEE Transactions on Speech and Audio Processing. 1994. V. 2. P. 291–298. doi: 10.1109/89.279278
4. Gemello R., Mana F., Scanzio S. et al. Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP. Toulouse, France, 2006. doi: 10.1109/ICASSP.2006.1660239
5. Seide F., Li G., Chen X., Yu D. Feature engineering in context-dependent deep neural networks for conversational speech transcription // Proc. IEEE workshop on Automatic Speech Recognition and Understanding, ASRU. Waikoloa, USA, 2011. P. 24–29. doi: 10.1109/ASRU.2011.6163899
6. Yao K., Yu D., Seide F. et al. Adaptation of context-dependent deep neural networks for automatic speech recognition // Proc. IEEE Spoken Language Technology Workshop. Miami, 2012. P. 366–369.
7. Liao H. Speaker adaptation of context dependent deep neural networks // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, 2013. P. 7947–7951.
8. Yu D., Yao K., Su H. et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, 2013. P. 7893–7897.
9. Li S., Lu X., Akita Y., Kawahara T. Ensemble speaker modeling using speaker adaptive training deep neural network for speaker adaptation // Proc. INTERSPEECH 2015. Dresden, Germany, 2015. P. 2892–2896.
10. Huang Z., Li J., Siniscalchi S.M. et al. Rapid adaptation for deep neural networks through multi-task learning // Proc. INTERSPEECH 2015. Dresden, Germany, 2015. P. 3625–3629.
11. Swietojanski P., Bell P., Renals S. Structured output layer with auxiliary targets for context-dependent acoustic modelling // Proc. INTERSPEECH 2015. Dresden, Germany, 2015. P. 3605–3609.
12. Price R., Iso K.-I., Shinoda K. Speaker adaptation of deep neural networks using a hierarchy of output layers // Proc. IEEE Workshop on Spoken Language Technology. South Lake Tahoe, USA, 2014. P. 153–158.
13. Karanasou P., Wang Y., Gales M.J., Woodland P.C. Adaptation of deep neural network acoustic models using factorised i-vectors // Proc. INTERSPEECH. Sungapore, 2014. P. 2180–2184.
14. Gupta V., Kenny P., Ouellet P., Stafylakis T. I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Florence, 2014. P. 6334–6338.
15. Senior A., Lopez-Moreno I. Improving DNN speaker independence with i-vector inputs // Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, 2014. P. 225–229. doi: 10.1109/ICASSP.2014.6853591
16. 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. P. 55–59. doi: 10.1109/ASRU.2013.6707705
17. Xue S., Abdel-Hamid O., Jiang H., Dai L., Liu Q. Fast adaptation of deep neural network based on discriminant codes for speech recognition // IEEE Transactions on Audio, Speech, and Language Processing. 2014. V. 22. N 12. P. 1713–1725.
18. Rath S.P., Povey D., Vesely K., Cernocky J. Improved feature processing for deep neural networks // Proc. INTERSPEECH. Lyon, France, 2013. P. 109–113.
19. Kanagawa H., Tachioka Y., Watanabe S., Ishii J. Feature-space structural MAPLR with regression tree-based multiple transformation matrices for DNN // Proc. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Hong Kong, 2015. P. 86–92.
20. Lei X., Lin H., Heigold G. Deep neural networks with auxiliary Gaussian mixture models for real-time speech recognition //

HMM-based speech recognition. *Computer Speech and Language*, 1998, vol. 12, no. 2, pp. 75–98.
3. Gauvain J.-L., Lee C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 1994, vol. 2, pp. 291–298. doi: 10.1109/89.279278
4. Gemello R., Mana F., Scanzio S. et al. Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*. Toulouse, France, 2006. doi: 10.1109/ICASSP.2006.1660239
5. Seide F., Li G., Chen X., Yu D. Feature engineering in context-dependent deep neural networks for conversational speech transcription. *Proc. IEEE workshop on Automatic Speech Recognition and Understanding, ASRU*. Waikoloa, USA, 2011, pp. 24–29. doi: 10.1109/ASRU.2011.6163899
6. Yao K., Yu D., Seide F. et al. Adaptation of context-dependent deep neural networks for automatic speech recognition. *Proc. IEEE Spoken Language Technology Workshop*. Miami, 2012, pp. 366–369.
7. Liao H. Speaker adaptation of context dependent deep neural networks. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Vancouver, Canada, 2013, pp. 7947–7951.
8. Yu D., Yao K., Su H. et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Vancouver, Canada, 2013, pp. 7893–7897.
9. Li S., Lu X., Akita Y., Kawahara T. Ensemble speaker modeling using speaker adaptive training deep neural network for speaker adaptation. *Proc. INTERSPEECH 2015*. Dresden, Germany, 2015, pp. 2892–2896.
10. Huang Z., Li J., Siniscalchi S. M. et al. Rapid adaptation for deep neural networks through multi-task learning. *Proc. INTERSPEECH 2015*. Dresden, Germany, 2015, pp. 3625–3629.
11. Swietojanski P., Bell P., Renals S. Structured output layer with auxiliary targets for context-dependent acoustic modelling. *Proc. INTERSPEECH 2015*. Dresden, Germany, 2015, pp. 3605–3609.
12. Price R., Iso K.-I., Shinoda K. Speaker adaptation of deep neural networks using a hierarchy of output layers. *Proc. IEEE Workshop on Spoken Language Technology*. South Lake Tahoe, USA, 2014, pp. 153–158.
13. Karanasou P., Wang Y., Gales M.J., Woodland P.C. Adaptation of deep neural network acoustic models using factorised i-vectors. *Proc. INTERSPEECH*. Sungapore, 2014, pp. 2180–2184.
14. Gupta V., Kenny P., Ouellet P., Stafylakis T. I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Florence, 2014, pp. 6334–6338.
15. 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
16. 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
17. Xue S., Abdel-Hamid O., Jiang H., Dai L., Liu Q. Fast adaptation of deep neural network based on discriminant codes for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2014, vol. 22, no. 12, pp. 1713–1725.
18. Rath S. P., Povey D., Vesely K., Cernocky J. Improved feature processing for deep neural networks. *Proc. INTERSPEECH*. Lyon, France, 2013, pp. 109–113.
19. Kanagawa H., Tachioka Y., Watanabe S., Ishii J. Feature-space structural MAPLR with regression tree-based multiple transformation matrices for DNN. *Proc. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit*

Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, 2013. P. 7634–7638.
21. Liu S., Sim K.C. On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Florence, 2014. P. 195–199.
22. Murali Karthick B., Kolhar P., Umesh S. Speaker adaptation of convolutional neural network using speaker specific subspace vectors of SGMM // Proc. INTERSPEECH 2015. Dresden, Germany, 2015. P. 1096–1100.
23. Tomashenko N.A., Khokhlov Y.Y. Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing // Proc. INTERSPEECH. Sungapore, 2014. P. 2997–3001.
24. Tomashenko N., Khokhlov Y. GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models // Proc. INTERSPEECH. Dresden, Germany, 2015. P. 2882–2886.
25. Tomashenko N., Khokhlov Y., Larcher A., Esteve Y. Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models // Lecture Notes in Computer Science. 2016. V. 9811. P. 304–311. doi: 10.1007/978-3-319-43958-7_36
26. Tomashenko N., Khokhlov Y., Esteve Y. On the use of Gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models // Proc. INTERSPEECH. San Francisco, USA, 2016. P. 3788–3792. doi: 10.21437/Interspeech.2016-1230
27. Tomashenko N., Khokhlov Y., Larcher A., Esteve Y. Exploration de paramètres acoustiques dérivés de GMMs pour l'adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds // Proc. 31reme Journrees d'Etudes sur la Parole (JEP), 2016. P. 337–345.
28. Tomashenko N., Khokhlov Y., Esteve Y. A new perspective on combining GMM and DNN frameworks for speaker adaptation // Lecture Notes in Computer Science. 2016. V. 9918. P. 120–132. doi: 10.1007/978-3-319-45925-7_10
29. Tomashenko N., Vythelingum K., Rousseau A., Esteve Y. LIUM ASR systems for the 2016 multi-genre broadcast arabic challenge // Proc. IEEE Workshop on Spoken Language Technology. San Diego, USA, 2016.
30. Rousseau A., Deleglise P., Esteve Y. Enhancing the TED-LIUM Corpus with selected data for language modeling and more TED talks // Proc. 9th Int. Conf. on Language Resources and Evaluation. Reykjavik, Iceland, 2014. P. 3936–3939.
31. Povey D., Ghoshal A., Boulianne G. et al. The Kaldi speech recognition // Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Waikoloa, USA, 2011. P. 1–4.
32. Матвеев Ю.Н. Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56. № 2. С. 47–51.
33. Povey D., Kanevsky D., Kingsbury B., Ramabhadran B., Saon, G., Visweswariah K. Boosted MMI for model and feature-space discriminative training // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Las Vegas, USA, 2008. P. 4057–4060. doi: 10.1109/ICASSP.2008.4518545
34. Tomashenko N.A., Khokhlov Y.Y. Fast algorithm for automatic alignment of speech and imperfect text data // Lecture Notes in Computer Science. 2013. V. 8113 LNAI. P. 146–153. doi: 10.1007/978-3-319-01931-4_20
35. Khokhlov Y., Tomashenko N. Speech recognition performance evaluation for LVCSR system // Proc. 14th Int. Conf. on Speech and Computer (SPECOM 2011). Kazan', Russia, 2011. P. 129–135.
36. Evermann G., Woodland P.C. Posterior probability decoding, confidence estimation and system combination // Proc. NIST Speech Transcription Workshop. 2000. V. 27. P. 78.
37. Maaten L.V.D., Hinton G. Visualizing data using t-SNE // Journal of Machine Learning Research. 2008. V. 9. P. 2579–2605.

and Conference. Hong Kong, 2015, pp. 86–92.
20. Lei X., Lin H., Heigold G. Deep neural networks with auxiliary Gaussian mixture models for real-time speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Vancouver, Canada, 2013, pp. 7634–7638.
21. Liu S., Sim K.C. On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Florence, 2014, pp. 195–199.
22. Murali Karthick B., Kolhar P., Umesh S. Speaker adaptation of convolutional neural network using speaker specific subspace vectors of SGMM. *Proc. INTERSPEECH 2015*. Dresden, Germany, 2015, pp. 1096–1100.
23. Tomashenko N.A., Khokhlov Y.Y. Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. *Proc. INTERSPEECH*. Sungapore, 2014, pp. 2997–3001.
24. Tomashenko N., Khokhlov Y. GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. *Proc. INTERSPEECH*. Dresden, Germany, 2015, pp. 2882–2886.
25. Tomashenko N., Khokhlov Y., Larcher A., Esteve Y. Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models. *Lecture Notes in Computer Science*, 2016, vol. 9811, pp. 304–311. doi: 10.1007/978-3-319-43958-7_36
26. Tomashenko N., Khokhlov Y., Esteve Y. On the use of Gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models. *Proc. INTERSPEECH*. San Francisco, USA, 2016, pp. 3788–3792. doi: 10.21437/Interspeech.2016-1230
27. Tomashenko N., Khokhlov Y., Larcher A., Esteve Y. Exploration de paramètres acoustiques dérivés de GMMs pour l'adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds. *Proc. 31reme Journrees d'Etudes sur la Parole (JEP)*, 2016, pp. 337–345.
28. Tomashenko N., Khokhlov Y., Esteve Y. A new perspective on combining GMM and DNN frameworks for speaker adaptation. *Lecture Notes in Computer Science*, 2016, vol. 9918, pp. 120–132. doi: 10.1007/978-3-319-45925-7_10
29. Tomashenko N., Vythelingum K., Rousseau A., Esteve Y. LIUM ASR systems for the 2016 multi-genre broadcast arabic challenge. *Proc. IEEE Workshop on Spoken Language Technology*. San Diego, USA, 2016.
30. Rousseau A., Deleglise P., Esteve Y. Enhancing the TED-LIUM Corpus with selected data for language modeling and more TED talks. *Proc. 9th Int. Conf. on Language Resources and Evaluation*. Reykjavik, Iceland, 2014, pp. 3936–3939.
31. Povey D., Ghoshal A., Boulianne G. et. al. The Kaldi speech recognition toolkit. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*. Waikoloa, USA, 2011, pp. 1–4.
32. Matveev Y.N. Study of informative speech features for automatic speaker identification. *Journal of Instrument Engineering*, 2013, vol. 56, no. 2, pp. 47–51. (In Russian)
33. Povey D., Kanevsky D., Kingsbury B., Ramabhadran B., Saon, G., Visweswariah K. Boosted MMI for model and feature-space discriminative training. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Las Vegas, USA, 2008, pp. 4057–4060. doi: 10.1109/ICASSP.2008.4518545
34. Tomashenko N., Khokhlov Y. Fast algorithm for automatic alignment of speech and imperfect text data. *Lecture Notes in Computer Science*, 2013, vol. 8113 LNAI, pp. 146–153. doi: 10.1007/978-3-319-01931-4_20
35. Khokhlov Y., Tomashenko N. Speech recognition performance evaluation for LVCSR system. *Proc. 14th Int. Conf. on Speech and Computer, SPECOM 2011*. Kazan', Russia, 2011, pp. 129–135.
36. Evermann G., Woodland P.C. Posterior probability decoding, confidence estimation and system combination. *Proc. NIST Speech Transcription Workshop*, 2000, vol. 27, pp. 78.
37. Maaten L.V.D., Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 2579–2605.

Авторы

Томашенко Наталья Александровна **–** аспирант, Лаборатория Информатики Университета Ле Мана (LIUM), Ле Ман, 72085, Франция; научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург 197101, Российская Федерация, tomashenko-n@speechpro.com

Хохлов Юрий Юрьевич **–** ведущий программист, ООО «ЦРТ-Инновации», Санкт-Петербург, 196084, Российская Федерация, khokhlov@speechpro.com

Ларшер Энтони **–** кандидат наук, доцент, Лаборатория Информатики Университета Ле Мана (LIUM), Ле Ман, 72085, Франция, anthony.larcher@univ-lemans.fr

Эстев Янник **–** доктор наук, профессор, директор, Лаборатория Информатики Университета Ле Мана (LIUM), Ле Ман, 72085, Франция, yannick.esteve@univ-lemans.fr

Матвеев Юрий Николаевич **–** доктор технических наук, главный научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; заведующий кафедрой, Университет ИТМО, Санкт-Петербург 197101, Российская Федерация; matveev@speechpro.com, matveev@mail.ifmo.ru

Authors

Natalia A. Tomashenko **–** postgraduate, Laboratory of Computer Science of the University of Le Mans (LIUM), Le Mans, 72085, France; researcher, “STC-Innovation”, Ltd., Saint Petersburg, 196084, Russian Federation; postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, tomashenko-n@speechpro.com

Yuri Yu. Khokhlov **–** leading programmer, "STC-Innovations", Ltd., Saint Petersburg, 196084, Russian Federation, khokhlov@speechpro.com

Anthony Larcher **–** PhD, Associate professor, Laboratory of Computer Science of the University of Le Mans (LIUM), Le Mans, 72085, France, anthony.larcher@univ-lemans.fr

Yannick Estève **–** D.Sc., Professor, Director, Laboratory of Computer Science of the University of Le Mans (LIUM), Le Mans, 72085, France, yannick.esteve@univ-lemans.fr

Yuri N. Matveev **–** D.Sc., Chief scientific researcher, “STC-Innovation”, Ltd., Saint Petersburg, 196084, Russian Federation; Head of Chair, ITMO University, Saint Petersburg, 197101, Russian Federation, matveev@speechpro.com, matveev@mail.ifmo.ru